DISSERTATION

ON

SPEAKER RECOGNITION

SUBMITTED FOR THE AWARD OF THE DEGREE OF

Master of Philosophy

IN

PHYSICS

Submitted by

PREETI VERMA

Under the supervision of

MR. S.K. GUPTA

DEPARTMENT OF PHYSICS
ALIGARH MUSLIM UNIVERSITY
ALIGARH-202002
(INDIA)

Mr. S.K. Gupta

Reader

DEPARTMENT OF PHYSICS
ALIGARH MUSLIM UNIVERSITY
ALIGARH – 202 002 (INDIA)
Phone : 0571-2701001 (O)
: +91-9412173531(mob)
Fax : 0571-2701001
E-mail : skg48@gmail.com

January 15, 2011

Certificate

This is to certify that ***Ms. Preeti Verma*** *has done her dissertation entitled* ***"Speaker Recognition"*** *in partial fulfillment for the award of the degree of Master of Philosophy in Physics and is based on the original work carried out by her under my supervision.*

(***Mr. S. K. Gupta***)

DEDICATED

TO

MY FAMILY

ACKNOWLEDGEMENT

First, I would like to acknowledge the grace of the Devine Providence who has encouraged me in numerous ways and whose benign benediction granted me the courage, patience and strength to embark upon this work and carry it to the successful completion.

I would like to express profound gratitude to my supervisor **Mr. S. K. Gupta**, Reader, Department of Physics, Aligarh Muslim University, Aligarh, for his invaluable support, encouragement and useful suggestions throughout this work. I have been amazingly fortunate to have an advisor who gave me freedom to explore by my own.

I am immensely grateful to **Prof. Wasi Haider**, Chairman, Department of Physics, Aligarh Muslim University, Aligarh, for providing me all the possible facilities available in the department in carrying out this work and encouragement in pursuit of course of study.

I am also very much thankful to **Dr. Omar Farooq**, Department of Electronics Engineering, Aligarh Muslim University, Aligarh, who has always empowered me with the knowledge and acted as a mentor and guided thereby contributing towards my professional growth. The completion of this dissertation would have been rather impossible without his wholehearted support and encouragement.

Special thanks to **Dr. Israr Khan**, Department of Physics, Aligarh Muslim University, Aligarh, for his continuous encouragement and motivation in one way or the other.

I owe my greatest debts to my family. I thank my parents for life and the strength and determination to live it. It is also because of them that I am able to complete this endeavor. Special thanks to my Grandparents for their love and support throughout my life. I also wish to thank my brother Mr. Shivam for sharing dreams with me and making me believe that I can achieve this.

I would like to express my special thanks to Ms. Sabiha Tabbasuum, Mr. Naseef Mohammad and Mrs. Naushaba Sami. I feel shortage of words to pay thanks to them who have been actively involved in offering suggestions or reviving on my dissertation.

I am also very much grateful to all my friends Ms. Anupam Sharma, Ms. Bhavana Gupta, Ms. Zoha Zainab, Ms. Anu Arora, Ms. Swapnil, Ms. Ummatul Fatima and Mr. Abdurahman. They have helped me stay same through these difficult years. Their support and care helped me overcome setbacks and stay focused on my goal. I greatly value their friendship and I deeply appreciate their belief in me.

Moreover I also express my appreciation to my seniors Dr. R. P. Sharma and Mrs. Jyoti Garg, who helped me to clear my doubts in doing this dissertation.

It's my pleasure to express my thanks to all the teaching and non-teaching staff and also to the seminar library staff of the Department of Physics, Aligarh Muslim University, Aligarh for providing me books, journals and other facilities.

Finally, I express my indebtedness to my glorious and esteemed institution, Aligarh Muslim University, Aligarh and U.G.C, for providing me financial assistance in the form of scholarship during my research programme.

Preeti Verma

(Preeti Verma)

CONTENTS

CHAPTER 1

INTRODUCTION

Speech is one of the most convenient means of communication between people. In speech a wealth of human experience, thought and emotion is conveyed by a series of gestures far more elaborate, stylized and quicker than any other human activity perhaps thought alone. The gestures are those of the tongue, lips, palate, larynx and the lungs just like when we write, the gestures are those of hand and fingers [1].The primary type of the spoken information is the word, which the speaker tries to pass to the listener. Beyond the unique human ability to receive and decode human language, the ear supplies us the ability to perform many diverse functions, for example, localization of objects, linguistic intelligence, enjoyment of music, speaker's identity, gender and many more.

Human is the only creature of nature whom she has endowed the power of speech. Joel Davis, the linguistic, states that, "Our brains are uniquely endowed with an innate ability to detect the basic rhythms and structures in sound or movement that can become the building blocks of symbolic communication [2]."

Communicating and understanding is central to human social life. As from the first stage of our life, when we do not know anything about how to communicate, then for expressing our thoughts we make babbling sounds and in doing this we get pleasure but these sounds cannot be called language as it does not form a social communication process. But as we start to learn arbitrary symbolic functions of words and our vocal sounds begin to acquire value, our mind undergo adjustment and we become integrated with society and now we can communicate with others with words.

To express our ideas, we convert them into a stream of suitable words arranged according to grammatical rules of the language. When we speak, our sound reaches our own ear as a feedback which helps us in controlling our articulatory organs to produce the desired words at appropriate intensity. Thus it is a type of close loop system and so it is often said that we speak with our ears. We can listen without speaking but we cannot speak without listening. This is the reason why, the people who are born without hearing ability learn to talk only with the greatest difficulty and a very few of them get success in producing that most of us would call normal speech [1].

According to the acoustic study, speech is a series of sound signals. Any signal is a physical quantity that varies with time or any other independent variable and is generated when a stimulus is exposed to a system. In case of speech signal, system is vocal cords and speech signal is generated by forcing air through vocal cords. The stimulus in combination with the system is called signal source. The information which is transferred through speech is mainly contained in sound waves and its fundamental frequency, intensity, etc. Such sound waves radiate through mouth and nostrils into the air, then these speech waves cause the eardrum to vibrate and it in turn passes the vibrations through the chain of small bones of the middle ear to the oval window of the inner ear. Thus by means of a marvelous mechanism the vibration energy is here transformed into nerve impulses which are sent to the brain through auditory nerve for interpretation.

To extend the man's capabilities and to increase the productivity of human beings, utilization of speech for communication between man and machine has been a significant requisite and the main motivative factor for developing speech interactive systems. The basic mechanism of speech communication with the machine is to functionally duplicate the behavior of human communication link, that is, the mechanism of speaking and understanding.

Today human communication, whether face to face, over the telephone network or through any other kind of audio-visual means is almost dominated by spoken language. Man's ever increasing tendency to use the speech is due to its faster information entry capability than other physiological means. Speech is a complicated signal produced as a result of several transformations occurring at several different levels: semantic, linguistic, articulatory and acoustic. Differences in these transformations appear as differences in the acoustic property of the speech signal. The speech signal processing is a diverse field with many applications as shown in the Figure 1.1.

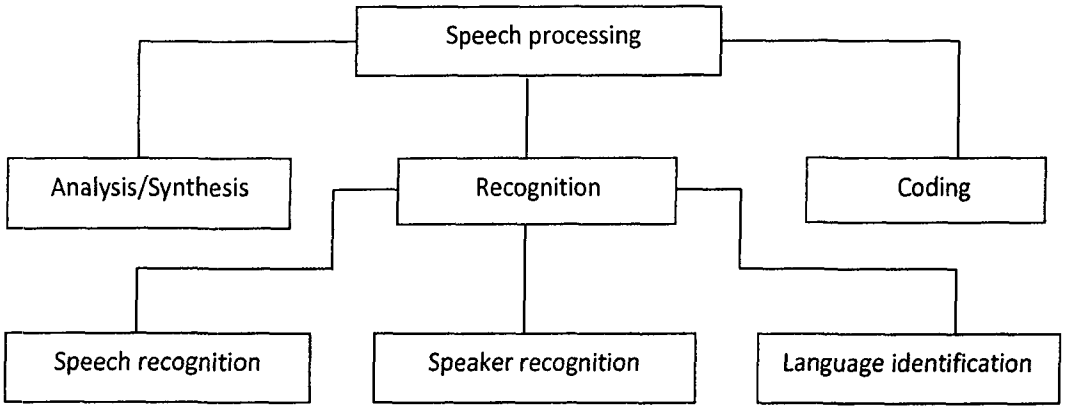


Figure 1.1: Applications of speech processing in various fields.

The speech signal conveys several levels of information, which can be taken out by performing different operations on it, as listed above. First, the area of speech recognition is concerned with extracting the underlying linguistic message in an utterance, i.e., understanding of spoken words. Second, the area of speaker recognition is concerned with extracting the identity of the person speaking the utterance. Third, although there are many languages spoken in different parts of the earth and each language has a system of speech sounds of its own, we can see a great similarity among these fundamental speech sounds. Also, the mechanism of producing particular speech sounds in the various languages is somewhat different, but the general mechanism is similar for all people.

Here we are interested in the field of speaker recognition. The area of speaker recognition attempts to recognize a speaker by his/her voice through measurements of the specifically individual characteristics arising in a speaker's voice. The idea is to identify the inherent differences in the articulatory organs (the structure of the vocal tract, the size of the nasal cavity and the vocal cord characteristics) and the manner of speaking. The branch of speech recognition is the task of understanding what is being said while the branch of speaker recognition is related with who is speaking. This task of speaker recognition is most useful when security applications through speech input are needed.

The human usually recognize the words of the speech after hearing and at the same time, they also recognize the characteristics of the speaker who is speaking those words. About 2-3 seconds of speech is sufficient to identify a familiar voice, although performance decreases for unfamiliar voices. Performance falls about 54% when duration is shorter than 1 second or distorted. Recognition often falls sharply when speakers attempt to disguise their voices, e.g., 59-81% accuracy depending on the disguise, vs. 92% for normal voices. This is reflected in machines, where accuracy decreases when mimics act as impostor. Human appears to handle mimics better than machine do. If the target (intended) voice is similar to the listener, he often associates the mimic voice with it. Certain voices are more easily mimicked than others, which tend further evidence to the theory that different acoustic cues are used to distinguish different voices [3].

From a general viewpoint, we can consider speaker recognition in three categories: recognition by humans (auditory speaker recognition), recognition by machine (automatic speaker recognition) and a compromise between the auditory and automatic recognition (semi-automatic speaker recognition).

The task of auditory speaker recognition is performed in our daily life. Naturally, we know that when we have heard a lot of speech from any person (may be our friend or our relative or any other known person); we can easily recognize his/her voice. The application of this task can be seen in forensic sciences, as the case of ear witness (the

person who heard the voice of the criminal during the crime). But, as the time between listening of the two voices increases, human performance decreases. Also, there are considerable differences between individuals in the auditory speaker recognition task. These are the arguments why, this method is not generally considered as a reliable method from a forensic point of view.

In case of speaker recognition, human and machine performance has been compared several times. It was observed from the analysis that individual human listeners vary significantly in their ability to recognize speakers on comparing the results with the state-of-the-art computer algorithms. Finally, it was concluded that human performs better when the quality of the speech samples is degraded. Though, since then, performance of computer algorithms has been improved significantly and the comparison may be already outdated.

The task of semi automatic speaker recognition performs both auditory (listening) and visual information (spectrogram, waveform). Two speech samples that are compared must be comparable in respect to their linguistic parameters. The selection of the units to be compared, e.g., phonemes or words must be carefully carried out. This requires an expert phonetician to segment the speech samples by hand. Also, an analysis to try recognizing which two of the spectrograms belong to the same speaker is performed. This example gives an idea about the complexity related to voice comparison [4].

Automatic Speaker Recognition is one area of artificial intelligence, where machine performance can exceeds human performance using short sentences and a large number of speakers and machine accuracy often exceeds that of human. This is especially true for unfamiliar humans, where the training time for speaker to learn a new voice well is very long compared with that for machine.

Acoustic Features Related to Speaker Recognition:

Speech data contains different types of information that convey speaker identity. The purpose of feature extraction stage is to extract the speaker specific information in the form of feature vectors at reduced data rate. These feature vectors include the information due to mainly one or more of the following: vocal tract, excitation source and behavioral traits. The speech signal is produced from the vocal tract system by varying its dimension with the help of articulators and exciting with a time varying source of excitation. The physical structure and dimension of the vocal tract as well as of the excitation are unique for each speaker. Also the behavioral traits like how the vocal tract and excitation source are controlled in speech production are unique for each speaker. This uniqueness is embedded into the speech signal and is used in speaker recognition. A good feature set should have the representation due to all the components of speaker information. To develop such a good feature set, it is necessary to understand the different feature extraction techniques developed so far.

In 1960, the study on digit recognition was conducted by P. Denes et al. suggested that the inter speaker differences exist in the spectral patterns of speakers [5]. This study motivated S. Pruzansky and then the first speaker identification study was conducted in 1963 using the spectral energy patterns as features. It was shown that the spectral energy patterns yielded the good performance, confirming their usefulness for speaker recognition [6]. Further a study was reported using the analysis of variance in 1964, by selecting a subset of features using F-ratio test defined as the ratio of variance of each speaker's feature distribution to the average of the variance of each distribution [7]. Speaker verification study was first conducted by Li et al. in 1966 using adaptive linear threshold elements [8]. This study used spectral representation of the input speech, obtained from a bank of 15 bandpass filters spanning the frequency range 300-4000 Hz. Two stages of adaptive linear threshold elements were trained from fixed speech utterances and operated on the rectified and smoothed filter outputs. A set of weights for the various frequency bands and time segments were resulted from the training process that

characterizes the speaker. Thus this study suggested that the spectral band energies contain the speaker information. Glenn et al. in 1967 suggested that the acoustic parameters produced by the nasal phonation are highly effective for speaker recognition [9]. In this study, average power spectra of nasal phonation were used as the feature for recognition. In 1969 [10], Fast Fourier Transform (FFT) based cepstral features were used in the speaker verification study. In this work, a 34-dimensional vector was extracted from speech data and it seems to provide a good representation of the speaker.

Most of the above studies used spectral patterns of speech as features for speaker recognition. In 1972 [11], Atal suggested the use of variations in pitch as a feature for speaker recognition. In addition to the importance of variations in pitch, Wolf in 1972, proposed other acoustic parameters such as, glottal source spectrum slope, word duration and voice onset time as important features for speaker recognition purpose [12]. Further, Atal introduced the concept of linear prediction in 1974 and suggested that Linear Prediction Cepstral Coefficients (LPCCs) are better than Linear Prediction Coefficient (LPC) and other features such as, pitch and intensity [13]. In general, the advantage of cepstral coefficients is that they can be derived from a set of parameters which are invariant to any frequency response distortion introduced by the recording or the transmission system.

Earlier studies neglected the features such as formant bandwidth, glottal source poles and higher formant frequencies, due to non-availability of measurement techniques. Thus the studies introduced after the linear prediction analysis explored the speaker specific potentials of these features for speaker recognition. But, a study reported by Rosenberg and Sambur suggested that adjacent cepstral coefficients are highly correlated and hence all coefficients are not necessary for speaker recognition [14]. In 1977, long term parameters averaging, which includes pitch, gain and reflection coefficients were studied for speaker recognition [15]. The reflection coefficients were highly informative and effective for speaker specific study. In 1985 [16], a study by G. R. Doddington suggested an approach for speaker verification that was different from the previous

approaches. Here the filter bank was not used but the speech was directly converted to pitch, intensity and formant frequency values, all sampled 100 times per second. Further, Furui introduced the concept of dynamic feature in 1986 to track the temporal variability in the feature vector for improving the speaker recognition performance [17].

A study by Reynolds in 1994 compared the different features like Mel Frequency Cepstral Coefficients (MFCCs), Linear Frequency Cepstral Coefficients (LFCCs), LPCCs and Perceptual Linear Prediction Cepstral Coefficients (PLPCCs) for speaker recognition [18] and finally suggested that MFCCs and LPCCs gave better performance than the other features. Though the MFCCs and LPCCs are used to extract the same vocal tract information, in practice these features differ in their performance due to the different principles involved in extracting it. Most of the studies discussed above considered vocal tract information as the speaker characteristics for speaker recognition. In 1995 [19], it was reported that Linear Prediction (LP) residual also contain speaker specific source information that can be used for speaker recognition. Thus it was suggested that though the energy of the LP residual alone gives less performance, combining it with LPCC improves performance as compared to that of the LPCC alone.

In 1999 [20], Chai Wutiwiwatchai suggested that apart from appropriate features, performance of the speaker recognition system directly depends on what pattern matching strategy is used. A text dependent speaker identification study was applied to Thai language using LPCs as feature vectors. This paper reported that Dynamic Time Warping (DTW) and Artificial Neural Network (ANN) are efficient for dependent task, while Vector Quantization (VQ) and Hidden Markov Model (HMM) are often used for text independent task. In 2000 [21], a study on speaker recognition was done taking only vowel phonemes as speech events. Among transformations of LPC parameters, the Adaptive Component Weighted (ACW) Cepstrum was shown to less susceptible to channel effects than others and was used as feature vector in this paper. The usefulness of residual signal has already been given in 1995. Further in 2002 [22], it was reported that improvement can be better if residual signal is computed through non-linear predictive

neural nets based model rather than linear prediction analysis. The study on speaker recognition was conducted in 2002 by using a new method, called RSFE method [23]. This method was used to discard the unreliable feature parameters and enhance the role of reliable parameters in the process of recognition with the help of weight coefficient. In 2004 [24], a new approach was suggested for improving the performance of a speaker identification system. In this approach, the speech signal was decomposed into various frequency bands using the multi-resolution property of the wavelet transform and then LPCCs and MFCCs were computed from each band. Finally, multiband approach was reported more effective and robust than the full band approach.

Furthermore, the work was carried out by keeping in mind the techniques that may reduce the effect of noise to improve the recognition performance. In 2005 [25], the logarithmic transformation in the standard MFCC analysis was replaced by the combined function of feature extraction process and speech enhancement methods to suppress the effect of noise. Another method was used for speaker recognition in noise environment in 2008 by applying the spectral subtraction process to assist the voice activity detection process [26]. In this paper, an energy based frame selection method was proposed for speaker modeling based on the spectral subtraction signal. In 2009, a new feature Weighted Mel Cepstrum (WMCEP) coefficient was used for speaker recognition purpose [27]. To obtain the WMCEP features, the psychologically weighted technology was applied in mel-cepstrum analysis using the Signal-to-Mask (SMR) as the weighting function. It was reported that these features not only describe the speaker's formants much better, but also had robustness to some extent for speaker recognition. Using the Wavelet based MFCCs in 2009 [28], Malik et al. suggested that after using the multi-resolution property of wavelet transform, it is better to use only the approximation coefficients to compute MFCCs. Here a study was also performed on the analysis for choosing the appropriate number of MFCCs, number of decomposition levels and wavelet type.

Most of the studies discussed so far have not considered features representing the behavioral traits for speaker recognition purpose. There are, of course, many different sources of speaker identifying information, including "high-level" information such as dialect, subject matter or context and style of speech (including lexical and syntactical patterns of usage). This high-level information is certainly valuable as an aid to recognition of speakers by human listeners, but it has not been used in automatic recognition systems because of practical difficulties in acquiring and using such information. Rather, automatic techniques focus on "low-level" acoustical features. These features include such characteristics of the speech signal as voice pitch frequency, formant frequencies and bandwidths and other factors such as, amplitude spectra of vowels and nasals and properties of the glottal source spectrum are also relevant.

Many researchers have done a lot of analysis on the feature extraction of speech useful for speaker characteristics. The essential goal is to find the non-linguistic information which is highly correlated with individual characteristics from the speech sounds. Usually when producing a speech sound, speaker's physiological and morphological features are encoded in acoustic characteristics of the sound. The diverse articulators contribute different physical properties in the acoustic spectrum that are personalized in the individual morphologies [29]. In order to extract that information some feature extraction techniques have been considered. Among these the mostly used are, in particular, MFCCs and LPCs. The main reason for the same may be the less intra-speaker variability and also availability of rich spectral analysis tools. Another relevant factor is the availability of straightforward techniques for improving the robustness of the features to channel distortions.

Motivation of the Research:

Speaker recognition is an example of biometric personal identification. This term is used to differentiate techniques that base identification on certain intrinsic characteristics of the person (such as voice, fingerprints, retinal patterns, or genetic structure) from those

that use artifacts for identification (such as keys, badges, magnetic cards, or memorized passwords), as it is well known that the intrinsic biometrics are presumed to be more reliable than artifacts, perhaps even unique. Thus a prime motivation for studying speaker recognition is to achieve more reliable personal identification.

For speaker recognition, the problem is how to extract and utilize the information that characterizes individual speakers. Generally, individual information of speaker results mainly from two factors: physiological and social factors. The first factor is related to the speaker's gender, age and oral morphology which are inborn characteristics. The other factor is concerned with the speaker's dialect, idiolect and occupation and so on, which results from his/her social environment. These factors derive from both the spectral envelope (vocal tract characteristics) and the supra-segmental features (voice source characteristics) of speech.

The main purpose of this dissertation is to start from understanding the speech production process, what is known about speaker individuality and then going into the details of feature extraction method, or in other words, we attempt to extract individual information that is involved in morphological details and acoustic characteristics and then implement it in speaker recognition. Although prosodic and other high level features have been exploited successfully in speaker recognition, our attention is on the low-level spectral features due to their text-independence, easy computation and widespread use.

In this dissertation, we focus on the problem of text-independent speaker identification using the best suited features LPCCs and MFCCs for classification. We have used the mean value of these features. It was reported that for the same phonemes extracted from different sentences but spoken by same speakers, the correlation of the mean value coefficient is high while for the same phoneme extracted from different speakers is different [30]. Thus this property characterizes the phonemes as well as the speakers. The one more advantage of using mean value is that the sentences can be used with their different lengths without any normalization. Finally, Linear Discriminant

Analysis (LDA) and multi layer neural network are used to classify the discriminative models. Also, Principal Component Analysis (PCA) method is applied to improve the recognition rate by reducing the dimensionality.

Previous Work on Automatic Speaker Recognition:

There is considerable speaker recognition activity in industry, national laboratories and universities. Automatic speaker recognition took off strongly in the 1960s. 'Visible speech', which used human examiners to recognize speakers from spectrograms output by analog computers, had been around since the 1940s (Potter, Kopp & Green) and continued as a research area to the end of the 1960s. But visible speech was not an automatic method for recognition. The Journal of Acoustical Society of America carried out most of the early work in both visible speech and automated talker recognition. The 1969 paper by Luck was the first to use cepstral coefficients as fundamental features in speaker recognition, as we have discussed earlier. This is highly significant because cepstral coefficients are still used today as the basis for many speech and speaker recognition systems. The point to be made here is that, at the most fundamental technical level biometrics is involving very slowly from its origin. The Table 1.1 represents the achievement in the area of speaker recognition. The following terms are used to define the columns: "source" refers to citation in the references, "features" are the signal measurements, "method" refers to the pattern matching process, "text" indicates the used mode of operation, "pop" is the population of the speakers used for database and error is the equal error percentage for speaker verification systems "v" or for speaker identification system "i". This table represents a simplified general view of past speaker recognition research.

Table 1.1: Speaker recognition progress

Source	Features	Method	Text	Pop	Error
James E. Luck [1969], [10]	First word length, pitch and cepstral vector	Distance classifier	Independent	26	v: 6.0% to 13.0%
Atal, [1974]	Cepstrum	Pattern Match	Dependent	10	i: 2.0% v: 2.0%
Markel and Davis, [1979]	LP	Long Term Statistics	Independent	17	i: 2.0%
Furui, [1981]	Normalized cepstrum	Pattern Match	Dependent	10	v: 0.2%
Schwartz, et al. [1982]	Log Area Ratios	Nonparametric pdf	Independent	21	i: 2.5%
Li and Wrench, [1983]	LP, Cepstrum	Pattern Match	Independent	11	i: 21.0% v: 4.0%
Doddington, [1985]	Filter-bank	DTW	Dependent	200	v: 0.8%
Soong, et al. [1985]	LP coefficients	VQ (size 64) Likelihood Ratio Distortion	10 isolated digits	100	i: 5.0% i: 1.5%
Higgins and Wohlford, [1986]	Cepstrum	DTW Likelihood scoring	Independent	11	v: 10% v: 4.5%
Attili, et al. [1988]	Cepstrum, LP, Autocorr	Projected Long Term Statistics	Dependent	90	v: 1.0%
Higgins, et al. [1991]	LAR, LP-Cepstrum	DTW Likelihood Scoring	Dependent	186	v: 1.7%
Tishby, [1991]	LP coefficients	HMM	10 isolated digits	100	v: 2.8% v: 0.8%
Reynolds, [1995]; Reynolds and Carlson, [1995]	Mel-Cepstrum	HMM (GMM)	Dependent	138	i: 0.8% v: 0.12%
Che and Lin, [1995]	Cepstrum	HMM	Dependent	138	i: 0.56% i: 0.14% v: 0.62%
Colombi, et al., [1996]	Cep, dCep, ddCep	HMM monophone	Dependent	138	i: 0.22% v: 0.28%

Reynolds [1996], [31]	Mel-Cepstrum, Mel-dCepstrum	HMM (GMM)	Independent	416	v: 11%/16% v: 6%/8% v: 3%/5% matched/mis-matched handset
Chai Wutiwiwatchai[1999], [20]	LPC	DTW	Dependent	20	i: 9.42%
Ehab F.M.F Bardan, [2000], [21]	Adaptive component weighted cepstrum	ANN	Dependent	10	v: 4.33% i: 7.0%
			Independent		v: 7.78% i: 11.05%
			Limited Vocabulary Recognition		2.5%
Macros Faúndez-Zanuy [2002], [22]	LPCC, residual	VQ (linear & non-linear codebook)	Independent	38	i: 3.68% (linear) i: 2.63% (non-linear)
Hassen Seddik [2004], [30]	Mean value of MFCC	Multilayer Neural Network	Independent	20	i: 23.0%
S.Malik, A.Afsar [2009], [28]	Wavelet based MFCC	VQ	Independent	64	i: 3.7% (non - telephonic) i: 13.23% (telephonic)

Overview:

This dissertation is divided into five chapters. Chapter 1 gives the introduction of the speaker recognition task and brief history about it. Chapter 2 discuss in detail the concepts and applications of speaker recognition purpose. Also the details about how humans perceive the speech and production methods, which can be used, based on the human auditory model to design the speaker recognition are given. The extensive details of feature extraction techniques used in this dissertation have been provided in chapter 3. This chapter also gives the details of classification techniques, such as LDA and neural network, which are used in the dissertation work for the design of speaker recognition

system. Furthermore, a dimensionality reduction technique PCA is also discussed. In chapter 4, the implementation of the techniques and various results obtained for used databases are given. In chapter 5, some techniques applied to improve the average recognition result obtained from the chapter 4 are discussed. Finally, the conclusion and future work is given.

CHAPTER-2

THEORY OF SPEAKER RECOGNITION

2.1 Concepts of Automatic Speaker Recognition:

There is an increasing need for person authentication in the world of information, applications ranging from credit card payments to border control and forensic sciences. In general, a person can be authenticated in two different ways:

One is, "Traditional authentication" method that includes something that the person has or knows, e.g., a key, credit card, PIN number or password. But the key or credit card can be stolen or lost while the PIN number or password can be easily misused or forgotten. The other way is, "Biometric person authentication" that includes person's signature, voice, fingerprints, facial features, etc. Each person has unique anatomy, physiology and learned habits that familiar persons use in everyday life to recognize the person. But there are various sources of errors in speaker recognition such as, intra-individual variation, voice disguise, mimicry and technical error sources.

The speaker's voice is not only affected by the physical and mental status of the speaker but also affected by stimulants and drugs. For instance, a chain-smoker has usually rougher voice than non-smoker. Since inter-session variability is probably the largest source of intra-speaker variation that is why, voice recorded even during the same day with the same technical conditions might not be matched correctly.

Second is, voice disguise means, if the speaker wants to change his/her voice knowingly so that it could not be matched with the another sample produced by the same speaker. Disguise is mainly common in forensic cases. For instance, when making a blackmail call, the criminal may keep his nostrils closed, talking with a pencil between

the front teeth, talking by whispering or by keeping a handkerchief on his mouth, or he might alter his voice during police investigation. On the other hand, mimicry is a special type of voice disguise where the speaker tends to map his voice to sound like another speaker. Disguise and mimicry definitely degrade the performance of a speaker recognition system.

Furthermore, there are several such sources that can degrade the performance of speaker recognition, both auditory and automatic. First, speech is recorded with a microphone or telephone handset and environmental noise (door slams, keyboard clicks, background babble, traffic noise, music, etc.) adds to the speech wave. Also, the poor quality microphone introduces non-linear distortion to the true speech. Speech coding can also degrade the speaker recognition performance significantly. But mismatched conditions are considered as the most serious error source in speaker recognition. It means that the circumstances for the training and testing phases are different.

The most common characterization is the division into two different tasks: "Speaker Identification" and "Speaker Verification".

Speaker identification task is to classify an unlabeled voice sample as belonging to (having been spoken by) one of a set of N reference speakers (N possible outcomes), or in other words, we can treat it as 1:N matching in which an unknown speaker is compared against a database of N known speakers and the best matching speaker is returned as the identified decision.

On the other hand, **speaker verification** task is to decide whether or not an unlabeled voice sample belongs to a specific reference speaker (two possible outcomes; the sample is either accepted as belonging to reference speaker or is rejected as belonging to an impostor), i.e., the verification task can be treated as 1:1 matching consists of making a decision when a given voice sample is compared against the claimed speaker. Thus an identity claim (e.g., a PIN code) is given to the system and the unknown speaker's voice sample is compared against the claimed speaker's voice template.

Of the identification and verification tasks, identification task is generally considered more difficult. This is intuitive when the number of registered speakers increases, the probability of incorrect decision increases. The performance of the verification task is not, at least in theory, affected by the population size since only two speakers are compared.

Speaker identification task is further classified into open-set and closed-set tasks. If the target speaker is assumed to be one of the registered speakers, the recognition task is closed-set problem, but if there is a possibility that the target speaker is none of the registered speakers, the task is termed as an open-set problem. In general, the open-set problem is much more challenging. In the closed-set task, the system makes a forced decision simply by choosing the best matching speaker from the speaker database, no matter how poor this speaker matches. However, in the case of open-set identification, the system must have a predefined tolerance level so that the similarity degree between the unknown speaker and the best matching speaker is within this tolerance. In this way, the verification task can be seen as a special case of the open-set identification task.

In speaker recognition, the first task is to distinguish between text-dependent methods and text-independent methods, for which the main distinction is as follows:

Text-dependent: The text to be spoken by the user is 'known' by the system.

Text-independent: There are no constraints on the text when the system is in use, so it must be trained to be able to cope with utterances of any text.

The feasibility of using text dependent methods will depend on the type of application and especially on whether or not the users can be regarded as being 'co-operative'. For example, text-dependent methods are appropriate for systems used in security applications, where speakers are expected to be co-operative and to know their own passwords. Also, sometimes for security applications, it is important to guard against attack by someone playing back a suitable voice recording of an authorized person. Text-

independent methods are therefore not really suitable for these applications. For surveillance and forensic applications, however, the speakers will most often not be co-operative (and may be ignorant of any surveillance), so it is not possible to use predetermined keywords and text-independent methods are required.

Text-dependent methods are also very vulnerable to attack if the user always speaks the same fixed text. This problem can be addressed by using a text-prompted method, whereby the user is prompted to enter a sequence of keywords that is chosen randomly every time the system is used. To reduce the risk of deception it is best to use sequences of words spoken in a naturally connected manner, so that it would be very difficult to record all possible combinations or to generate natural-sounding imitations even with a fairly sophisticated concatenation machine. The greater the variety of different sequences that may be requested, the less opportunity there is for the system to be deceived, but the greater the enrolment effect.

If the text is fixed, only one reference is required. An input sample is time aligned with this reference and a similarity measure (a distance for a template method, or a likelihood in the case of the HMM approach) is accumulated for the duration of the utterance. Fixed-text methods tend to give the best speaker recognition performance because the text is known and hence these methods can fully exploit the speaker individuality that is associated with each sound in the utterance. While text-prompted speaker recognition is more difficult because the system must recognize both the identity of the speaker and verifying the utterance is indeed a spoken version of the specified text. Thus an utterance should be rejected if its text differs from the prompted text, even if it is spoken by the registered speaker [32].

All technologies of speaker recognition: identification and verification, text dependent or independent, each has its own advantages and disadvantages and may require different treatments and techniques. All speaker recognition systems contain mainly two main modules: feature extraction and matching. The task of feature extraction

is concerned with finding a means of measuring the "distinguishability" of the distributions. It transforms the raw speech signal into a compact but effective representation by extracting a small amount of data from the speech signal that can later be used to represent each speaker. The quality of later components (speaker modeling and pattern matching) is strongly determined by the quality of feature extraction. In other words, classification can be at most as accurate as the features. Feature matching involves the actual procedure to identify the unknown speaker by comparing extracted features from his/her voice input with the ones from a set of known speakers.

Speech exhibits significant variation from instance to instance for the same speaker and text and a speaker produces a stream of features that characterize both the speech as well as the speaker. More than a few seconds of speech, we expect the features to fill feature space in a way that depends primarily on the speaker, not the particular text spoken. The assumption is that with sufficient speech, a good representation of the sounds that a speaker can create will be observed. The goal is to obtain descriptions or model of a speaker's pattern in feature space which can be used to identify the speaker of a test utterance.

An important step in speaker identification process is to extract sufficient information for good discrimination and at the same time, to have captured the information in a form and size that is amenable to effective modeling. The amount of data generated by short utterances is quite large. The process of reducing data while retaining classification information falls under the general heading of feature extraction. The vectors extracted are termed as features and the n-dimensional feature space is referred to as speaker space.

The distributions of different speakers overlap and share speaker space, but are ideally distinguishable from each other so that speaker identification can be achieved. Speaker identification is accomplished by determining how and where voices "spend their time" in speaker space. To simplify matter, it is usually assumed that the feature

vectors are independent of one another, even though vectors from consecutive frames are correlated in reality.

For speaker recognition, the feature analysis needs to capture the characteristics of different talkers. Ideally features should be chosen that maximize the separation between individuals, while not being too sensitive to occasion-to-occasion variation within the speech of one person. For some applications, the robustness to within speaker variation will need to include variability that may be introduced by an individual attempting to disguise his or her voice. There are many applications, such as that requiring speaker recognition over the telephone, for which the feature representation also needs to be robust to noise and to channel variations. Otherwise these variations could cause changes to the features that are larger than the differences between speakers.

In feature matching stage, all the speaker recognition systems have to serve two distinguish phases. The first one is referred to the enrollment session or the training phase while the second one is referred to as the operation session or the testing phase. In the training phase, the registered speaker has to provide samples of their speech so that the system can train a reference model for that speaker. In case of speaker verification system, a speaker specific threshold is also computed from the training samples. During the testing phase, the input speech is matched with stored reference model and recognition decision is made.

2.1.1 Applications:

Speaker recognition is a difficult task and it is still an active research area. It is based on the premise that a person's speech exhibits characteristics that are unique to the speaker. The main advantage of speaker recognition is its naturalness. Speech is our main communication matter and embedding speaker recognition technology into applications is non-intrusive from the user's viewpoint. Another advantage is cheap cost; no special equipment is needed. In order to capture a speech signal, only a microphone is needed, as contrast to fingerprint and retinal scanners, for instance. Signal processing and pattern

matching algorithms for speaker recognition are low cost and memory efficient, and thus applicable for mobile devices. Last but not the least, performance of speaker recognition is considerably high in right conditions [4].

For speaker recognition, fingerprints or iris analysis are good examples of other biometric approximations to person identification, where the test sample is directly matched with the known pattern. However, voice identification must be accomplished from a different point of view, in an analogous way to face recognition or graph logical analysis of hand writing, as signal variability (written signs, facial features or speech characteristics) incorporates to the identification process an additional level of complexity [33]. The speaker recognition technology enables access level of various sources by voice (Furui, 1991, 1997, 2000). Applicable services include the following:

- Transaction authentication: Toll fraud prevention, telephone credit card purchases, telephone brokerage (e.g., stock trading), information and reservation services.
- Access control: Physical facilities, computers and data networks.
- Monitoring: Remote line and attendance logging, home parole verification, prison telephone usage.
- Information retrieval: Customer information for call centers, audio indexing (speech skimming device).
- Knowledge authentication along with voice authentication.
- Home shopping that includes speaker identification and verification based on home phone number, provide secure access to customer record and credit card information.
- Another important application of this technology is as a forensic tool [34]. If there is a speech sample that was recorded during the commitment of a crime, the

suspect voice can be compared with this in order to give an indication of the similarity of the two voices.

2.2 Speech Production Process:

It has been conjectured that speech evolved when ancient man discovered that he could supplement his communicative hand signals with related gestures of his vocal tract. The speech production process begins from the speaker when he formulates a message in his mind that he wants to transmit to the listener via speech. Then it converts this message into a language code. Once the language code is chosen, the speaker must executes a series of neuromuscular commands to cause the vocal cords to vibrate when appropriate and to shape the vocal tract such that the proper sequence of speech sounds is created and spoken by the speaker, thereby producing an acoustic signal as the final output. The Figure 2.1 shows the proper sequence of speech production as explained above. Understandings of how human produce sounds form the basis of speaker recognition.

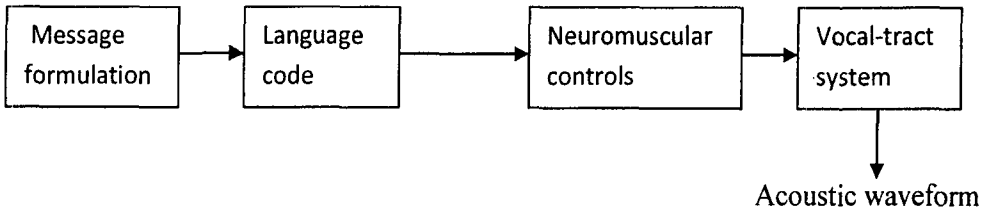


Figure 2.1: Proper sequence of speech production process [35].

2.2.1 Anatomy:

The vocal tract is generally considered as the speech production organs above the vocal cords. A schematic diagram of the human vocal mechanism is shown in Figure 2.2. The vocal tract includes the following:

Laryngeal pharynx (beneath the epiglottis),

Oral pharynx (behind the tongue, between the epiglottis and velum),
Oral cavity (towards the velum and bounded by the lips, tongue and palate),
Nasal pharynx (above the velum, rear end of the nasal cavity),
Nasal cavity (above the palate and extended from the pharynx to the nostrils).

The vocal cords are shown in the Figure 2.2. The larynx is composed of the vocal cords, the top of the cricoid cartilage, the arytenoid cartilages and the thyroid cartilage (also known as "Adam's apple"). The vocal cords are stretched between the thyroid cartilage and the arytenoids cartilages. The area between the vocal cords is called the glottis.

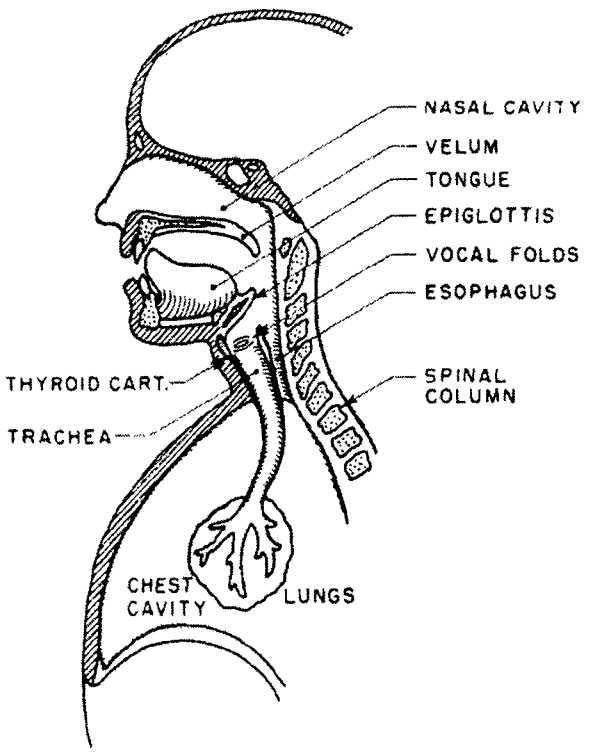


Figure 2.2: Schematic diagram of human vocal mechanism [35].

The vocal tract is an acoustical tube which is non uniform in cross sectional area. The Figure 2.3 shows a longitudinal cross section X-ray of human vocal apparatus. It begins at the opening of the vocal cords, or glottis and ends at the lips. In an average male, the length of the vocal tract is about 17 cm. Its cross-sectional area is determined by the positions of the tongue, lips, jaw and velum, varies from zero (in case of complete closure) to about 20 square cm. All these components are called articulators by speech scientists and move to different positions to produce various sounds. The nasal tract begins at the velum and ends at the nostrils. When the velum is lowered, the nasal tract is acoustically coupled to the vocal tract to produce the nasal sounds of speech. Acoustic coupling between nasal tract and vocal tract is controlled by the size of the opening at the velum.

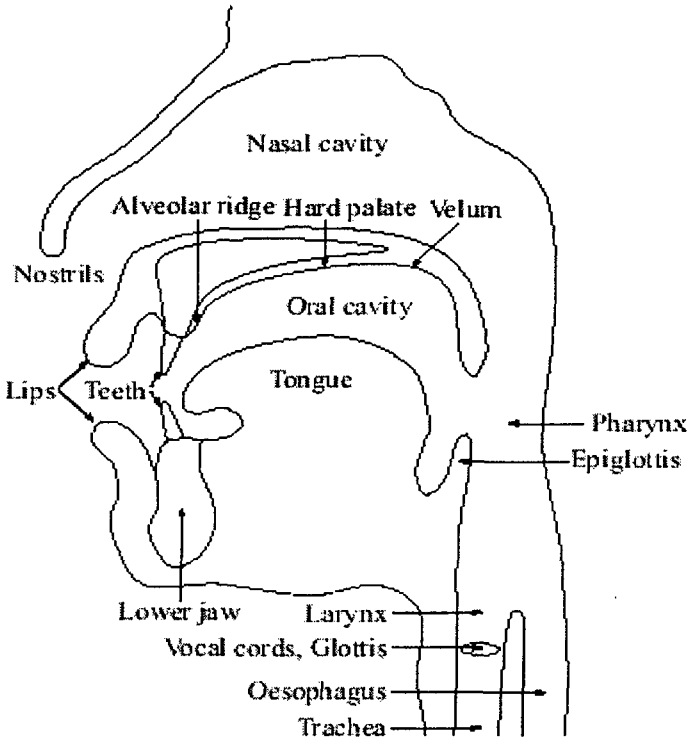


Figure 2.3: Longitudinal cross section X-ray of the human vocal apparatus [36].

From the technical point of view, it is more useful to think about speech production system in terms of an acoustic filtering operation that affects the air going from the lungs. There are three main cavities: nasal, oral and pharyngeal cavities that comprise the main acoustic filter. The articulators are responsible for changing the properties of the system and forming its output. Combination of these cavities and articulators is called vocal tract. Its simplified acoustic model is shown in Figure 2.4. Human vocal mechanism is driven by an excitation source. When we take breath, air enters into the lungs. The excitation is generated by airflow from the lungs carried by trachea (or "windpipe") through the vocal cords. The muscle force pushes air out of the lungs as a piston pushing up within a cylinder. The airflow causes the tensed vocal cords to vibrate and then this air flow is chopped into the quasi periodic pulses, which are modulated in frequency when it passes through pharynx (throat cavity), oral cavity and possibly the nasal cavity. Thus produced different sounds depend upon the various positions of articulators (i.e., jaw, tongue, velum, lips and mouth). The excitation can be characterized as phonation, whispering, frication, vibration or a combination of these.

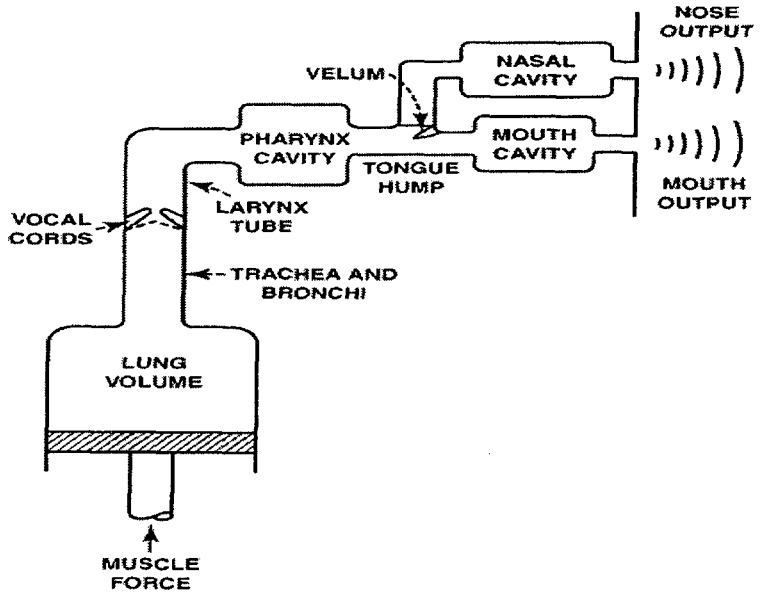


Figure 2.4: Simplified representation of human speech production [35].

The airflow causes the tensed vocal cords to blow apart by building up the pressure underneath them and hence, the vocal cords are drawn back together again by their tension and elasticity. This pulsed air stream, arising from the oscillating vocal cords, excites the vocal tract and speech sounds are produced. Production of sounds in this manner is called phonation and the speech sounds produced by phonated excitation are called voiced sounds. The frequency of oscillation is called the fundamental frequency, and it depends upon the length, tension and mass of the vocal cords.

Another source of excitation is produced by turbulent flow of air. When the vocal cords are relaxed, the airflow either must pass through a narrow constriction in the vocal tract, or a third source of excitation is created by building up the pressure at some point of closure. The sounds produced by turbulent airflow are called unvoiced sounds. While in the second case, an abrupt release of the pressure provides a transient excitation of the vocal tract causing a brief transient sound. Whispered excitation is produced by turbulent airflow through a small opening between the arytenoids cartilages at the partially closed vocal cords and has a wide band noise spectrum.

The sequence of sounds is called speech. According to the state of vocal cords, and positions of different articulators (jaw, lips, mouth, lungs, etc.), different speech sounds are produced. In speaker recognition task, we are interested in the physical properties of human vocal tract. In general, it is assumed that vocal tract carries most of the speaker related information [31]. All parts of human vocal tract described above can serve as speaker dependent characteristics and are called physical distinguishing factors.

2.2.2 Vocal Model:

In order to develop an automatic speaker recognition system, we should construct reasonable model of human speech production system. Having such a model, we can extract the properties from the signal and using them, we can decide whether or not these two signals belong to the same model and as a result to the same speaker.

Modeling process is usually divided into two parts: the excitation (or source) modeling and the vocal tract modeling. This approach is based on the assumption of independence of the source and the vocal tract models. Let us look first at the continuous time vocal tract model called multitube lossless model, which is based on the fact that production of speech is characterized by changing the vocal tract shape. Because the normalization of such a time varying vocal-tract shape model is quite complex, in practice it is simplified to the series of concatenated lossless acoustic tubes with varying cross-sectional areas. Tract model serves as a transition to the more general discrete-time model, also known as source-filter model, which is shown in Figure 2.5.

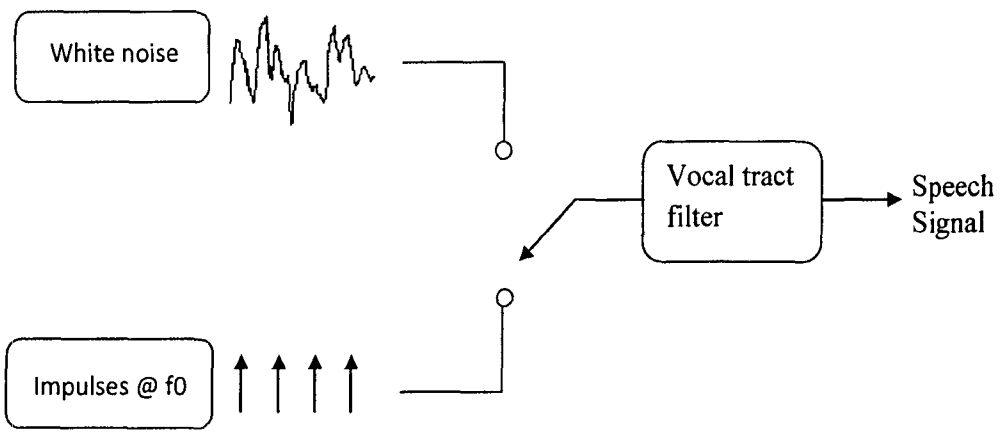


Figure 2.5: The source-filter model of speech production [37].

In this model, the voice source is either a periodic pulse stream or uncorrelated white noise, or a combination of these. This assumption is based on the evidence from human anatomy that all types of sounds, which can be produced by humans, are divided into three general categories: voiced, unvoiced and combination of these two. Finally, we can think about the vocal tract as a digital filter, which affect the source signal and about produced sound output as a filter output. Then based on the digital filter theory, we can extract the parameters of the system from the output.

The issues described in this chapter serve as the basis for developing speaker recognition techniques described in next chapter.

2.3 Speech Sounds and Features:

Speech sounds can be analyzed, described and classified from different point of view in relation to the chain of events that take place when the speaker communicates the message to the listener. Articulatory, acoustics and perceptual aspects of speech are normally considered for such an analysis. While human can produce an infinite number of sounds (within the constraints of the vocal tract), each language has a small set of abstract linguistic units called phonemes which represents the sounds of speech. Phonemes are the minimal constructive units of the language. Each language typically has about 40 phonemes [38], which provide an alphabet of sound.

2.4 Classification of Speech Sounds:

Speech sound classification is usually attained according to their manner and place of production. The alphabets of most of the major languages of the world in modern times are represented by the alphabet of International Phonetic Association (IPA). There are 10 vowels and 31 consonants in Hindi language [36] that are discussed below:

2.4.1 Vowels:

From the acoustic point of view, vowels are voiced, long in duration and have largest amplitude among phonemes. These are produced by exciting the essentially fixed vocal tract shape with quasi periodic pulses of air due to vibration of vocal cords. The manner by which the cross-sectional area along the vocal tract varies, determines the resonant frequencies of the vocal tract (formants) and the sound is produced.

The vowel produced is determined primarily in terms of the tongue hump position (i.e., front, mid, back) and tongue hump height (high, mid, low), but the positions of lips, mouth and possibly the velum make influence upon it. The tongue hump is the mass of tongue at its narrowest constriction within the vocal tract. The front vowel shows a pronounced high frequency resonance, the back vowel shows predominance low frequency spectral information and mid vowel shows a balance of energy over a broad frequency range. The classification of frequently used Hindi vowels by the tongue hump position and degree of constriction is shown in Table 2.1.

Table 2.1: Classification of Hindi vowels according to the tongue hump position and degree of constriction

Degree of constriction	**Tongue hump position**	
	Front	**Back**
High	/i/ /ɪ/	/u/ /ʊ/
Medium	/e/ /ɛ/	/o/ /ɔ/
Low	/ʌ/	/a/

2.4.2 Consonants:

The consonants constitute those sounds which are not exclusively voiced, mouth radiated and has a relatively stable vocal configuration. The consonant is a sound in a spoken language that is characterized by a closure sufficient to cause audible turbulence, at one or more points along the vocal tract. The word consonant comes from Latin meaning "sounding with" or "sounding together", the idea behind that consonants don't sound on their own, but only occur with a nearby vowel. In Hindi speech, we have 31 consonants in which 16 are stops, 3 fricatives, 4nasals, 2 glides, 4 affricates and 2 continuants as

given in Table 2.2. These sounds are determined according to how and which speech organs come into contact with one another and how the air is finally released from the mouth. An speech organ which influences the flow of air in the oral cavity is an 'articulator'. Thus the consonants are broadly classified according to manner and place of articulation, discussed in details below.

Manner of articulation:

This category of classification is made on the basis of how the vocal tract restricts the airflow. According to this, the consonants are divided into eight articulatory features given below:

1. **Fricative consonants:**

   Fricatives are produced from an incoherent noise excitation of the vocal tract. If the pressure behind the constriction is fast enough and the constriction is sufficiently narrow then airflow also becomes fast enough to generate the pressure at the end of the constriction. If the vocal cords are in conjunction with the noise source then the fricative is voiced and if only the noise source is associated, then the produced fricative is unvoiced.

2. **Nasal Consonants:**

   Nasal consonants are also called as murmur and sometimes nasal stops. These are excited by the vocal cords and hence are voiced. The vocal tract is completely constricted along the oral passage way and as the velum is lowered, the air flows through the nasal tract and the sound radiates at the nostrils. Although the oral cavity is constricted in front, it is acoustically coupled to the pharynx, thus the mouth serves as the resonant cavity that traps the acoustical energy at some natural frequencies.

3. **Stop consonants:**

These sounds are also called plosives. These are transient, non-continuant sounds, depend upon vocal tract dynamics and are produced when a complete closure is formed at some point in the vocal tract. The lungs build up the pressure behind this occlusion and the pressure is suddenly released by an abrupt motion of the articulators. The stops can be produced with or without simultaneous voicing. As stop consonants are dynamic in nature, these are influenced by the vowels which follow them.

4. **Continuants:**

These sounds /w/ and /j/ are called continuants as they represent a peculiar way of speaking means mouth is adjusted to say vowel /U/ and /I/ respectively and ends at any other vowels. All others are treated as non-continuants. These sounds are also treated as semivowels because the oral passage way is more constricted than in most vowels and the tongue tip is not down.

5. **Liquids:**

The sounds / l / and / r / that greatly resemble with vowels are called liquids. These are also called as lateral and trill respectively. These are characterized by vocal excitation of the tract, no effective nasal coupling or sound radiated from the mouth. They have spectral very similar to vowels, but are usually a few dB weaker.

6. **Affricates:**

As in dipthongs pair of vowels combined, these are produced by the combination of stops and fricatives.

7. **Voicing:**

   If the vocal cords vibrate during the production of sound, the sound is called voiced sound, otherwise it is unvoiced.

8. **Aspirated:**

   Aspiration is defined as glottal friction produced with (for voiced sounds) or without (for unvoiced sounds) glottal pulsing, while the glottis is narrowly or widely open and supraglottal vocal tract is unobstructed. 12 consonants in Hindi speech /p^h, t^h, $ʈ^h$, k^h, b^h, d^h, $ɖ^h$, g^h, h, $tʃ^h$, dz^h, r^h/ are treated as aspirated sounds. The sound that does not represent aspirated character is treated as unaspirated sound.

Place of articulation:

Place of articulation refers to the location or point of constriction made along the vocal tract by the articulators. This classification is made on the basis of which speech organ obstruct the airflow. According to the place of articulation, there are nine categories of consonants as:

1. **Bilabial:** The upper and lower lip clamp together for a moment stopping the air before releasing it.
2. **Labio-dental:** The upper teeth come in contact with the lower lip.
3. **Dental:** The two rows of the teeth and tongue touch each other for a brief period.
4. **Alveolar:** The tip of the tongue touches the alveolar ridge obstructing the air.
5. **Post-alveolar:** The tip of the tongue touches the region behind the alveolar ridge.
6. **Palatal/Palatal-alveolar:** The middle of the tongue touches hard palate or the tip the tongue touches the region past alveolar and before the palatal.
7. **Retroflex:** The tongue tip is curled up and back.

8. **Velar:** Back of the tongue touches the soft palate.

9. **Glottal:** The obstruction of air takes place between the vocal cords.

Table 2.2: Classification of Hindi stop consonants

	Bilabial	**Dental**	**Retroflex**	**Palatal**	**Velar**	***Glottal***
Fricatives		/s/		/ʃ/		/h/
Stops	/p,pʰ/ /b,bʰ/	/t,tʰ/ /d,dʰ/	/ʈ,ʈʰ//ɖ,ɖʰ/		/k,kʰ//g,gʰ/	
Affricates				/tʃ,tʃʰ/ /dz,dzʰ/		
Lateral		/l/				
Continuants	/w/			/j/		
Nasals	/m/	/n/	/ñ/		/ŋ/	
Trill			/r/			

In the table, the one on the right is a voiced, while on the left is unvoiced consonant.

2.5 Sound spectrograms

Since sound is a transitory phenomenon and is to be perceived subjectively by listeners only, one use methods of making the sound permanent by displaying it in visual form. The commonest visual display of an acoustic signal is called the sound spectrogram. To make a spectrogram, a Fourier transform is applied to an acoustic wave, deriving the frequencies and amplitudes of its component waves. Depending on the size of the Fourier analysis window, different levels of resolution are achieved.

A spectra or spectrum is a two dimensional representation. In spectrum, frequency is along horizontal axis and intensity is along vertical axis. For a section of sound wave, spectrum shows what the frequency components are present and what the relative intensities are. Spectrum has no time representation and therefore, provides a static picture while spectrogram gives dynamic picture of the stretch of a sound. It shows the

change of acoustic parameters over time. Spectra are helpful in an utterance, where more precise amplitude-frequency measurements are desired at a certain point.

There are two main kinds of voice analysis performed by the spectrograph, wideband (with a bandwidth of 300-500 Hz) and narrowband (with a bandwidth of 45-50 Hz). In a wideband spectrogram, a small window is used and hence adjacent harmonics are smeared together and time resolution is better. Thus makes broader distinction in frequency and finer distinction in time. For narrow band spectrogram, a long time window is used that reveals harmonics and hence resolves frequency at expense of time. Thus makes narrower distinction in frequency but broader distinction in time

The Figure 2.7(a-c) shows the waveform, wideband spectrogram and narrowband spectrogram for the word 'gazi' taken from the sentence of ELSDSR database. In general, the darker is the color, the greater is the amplitude of that particular frequency component. Voice researchers use the spectrograph as a tool for analyzing vocal output.

(a)

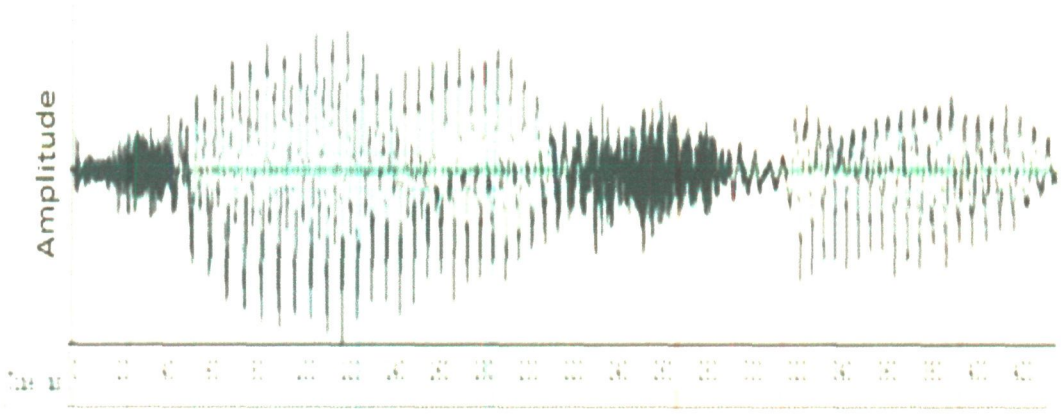
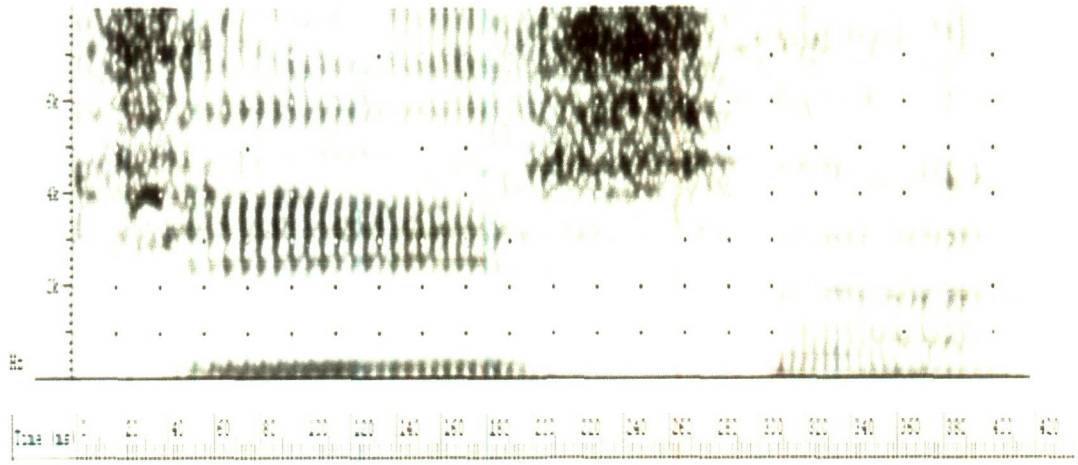


Figure 2.7: (a) waveform for the word /gazi/ using 'sfs' software.

(b)



(c)

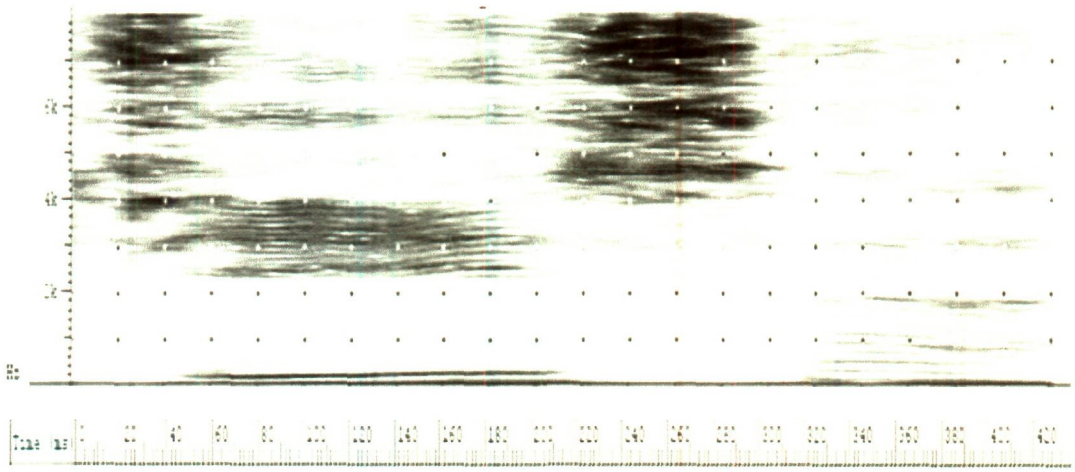


Figure 2.7: (b) wide band spectrogram (c) narrow band spectrogram for the word /gazi/ using 'sfs' software.

CHAPTER 3

FEATURE EXTRACTION AND SPEAKER MODELLING

3.1 Phases of Automatic Speaker recognition

Much of the theory of speaker identification is common to speaker verification. Like most problems, speaker recognition may also be considered to be divided into three parts: measurement, feature extraction and classification [39], as shown in the Figure 3.1, given below.

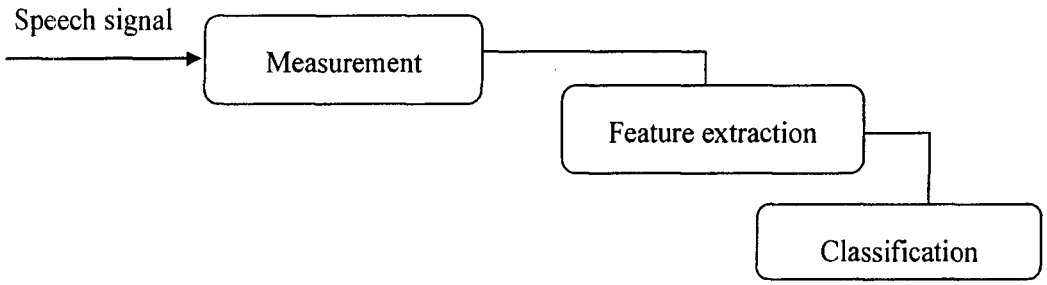


Figure 3.1: Stages of automatic speaker recognition [39].

The preprocessing of measurement part is necessary and common to all speech processing applications. In this part, we work with sampling, framing, windowing and so on. The task of this preprocessing stage is to extract all relevant acoustic information in a compact form compatible with the acoustic models to suit the speaker recognition system and other applications.

The process of feature extraction is to convert the speech signal into a sequence of feature vectors carrying characteristic information to be used to classify the signal. The extraction of salient feature vectors is a key step in solving any pattern recognition problem. These vectors are used as the basis for various types of speech analysis algorithms. It is typical for such algorithms to be based on features computed on a window basis. These window based features can be considered as short time description of the signal for that particular moment in time. The performances of a set of features depend on the application to be used.

In the third stage, i.e. the classifier takes the features computed by the feature extractor and performs either template matching or probabilistic likelihood computation on the features depending on the type of algorithm applied. Before it can be used for classification, the classifier has to be trained so that a mapping from the feature to the label of a particular class is established. Each stage is discussed in details in the next section.

3.2 Measurement Analysis:

The measurement and feature extraction stage normalize the collected data and transform them to the feature space. During the past few years, pre-processing of text independent speaker recognition has become the normalization formula. This stage is common for all feature extraction techniques by processing further the raw digital signal in order to improve the performance of the recognition system, or to prepare the speech for feature calculation stage. The main steps in pre-processing are given in details below [40].

3.2.1 Pre-emphasis:

Usually speech is pre-emphasized before any further processing. Pre-emphasis is a process of passing the signal through a filter, which emphasizes higher frequencies. In speech signal, the most part of the energy is carried by the low frequencies. When the frequency increases, pre-emphasis also increases the energy of the signal. It also serves to

emphasis the formant peaks, to make them more "visible" in the spectrum. Pre-emphasis makes the spectrum more flat by raising the energy in high frequency region. In digital signal processing system, pre-emphasis is a digital high pass filter which processes the digitized speech signal. The high-pass filtering action may be achieved digitally using the first order difference equation. Such filter is implemented by the formula:

$$y[n] = x[n] - ax[n-1], \tag{3.1}$$

where, $y[n]$ denotes the current output sample (n is the sample index) of the pre-emphasis filter, $x[n]$ is the current input sample, $x[n-1]$ is the previous input sample and 'a' is a pre-emphasis constant and its transfer function is given by:

$$H(z) = 1 - az^{-1} \tag{3.2}$$

where, $a > 0$ controls the slope of the filter. In speech recognition experiments, 'a' varies from 0.94 to 0.97.

It is also noted that it doesn't really matter if pre-emphasis done before or after windowing.

3.2.2 Normalization and Mean Subtraction:

Due to possible mismatch between training and testing conditions, it is considered in practice to reduce the amount of variation in the data that does not carry important speech information as much as possible. For instance, differences in loudness between recordings are irrelevant for recognition. For reduction of such irrelevant sources of variation, normalization transforms are applied. In the normalization process, every sample value of the speech signal is divided by the highest amplitude sample value. For removing the DC offset and some of the disturbances introduced by recording instruments mean value of the speech signal is subtracted from each sample.

3.2.3 Framing and Windowing:

Speech is a non-stationary signal and its characteristics may change in very short time instances. In order to capture the variability in the waveform, every state of the art system firstly segments the signal into frames. At a given, very short time frame, the speech segment is close to stationary. In this interval speech signal remain unchanged. The features are extracted from these frames. The original Fourier transform operates on a signal of theoretically infinite length, so the short time frame analysis (STFT) requires that each frame somehow explain to infinite length. For extracting the spectral features of a speech signal a short-time analysis is applied. Once the frame blocking procedure is completed, to every frame a windowing function is applied to suppress the effect of discontinuities at frames edges. The purpose of the windowing is to reduce the effect of the spectral artifacts that result from the framing process. Windowing in the time domain is a point-wise multiplication of the frame and the window function. According to the convolution theorem, this corresponds to convolution of the short-term spectrum with the window function magnitude response. The most popular window is the Hamming window given by:

$$w[n] = \begin{cases} 0.54 - 0.46\cos\frac{2\pi n}{N} & ,0 \leq n \leq N \\ 0 & ,otherwise \end{cases} \tag{3.3}$$

Hamming window is applied on each frame to minimize the signal discontinuities at the beginning and end of the each frame during truncation of the signal. When these windowing functions are applied to a signal, some information near the frame boundaries is lost. For this reason, a further improvement to the STFT is to overlap the frame with 30-50%. When each part of the signal is analyzed in more than one frame, information that is lost at a frame boundary is picked up between the boundaries of the next frame. The overlap between two consecutive frames is also necessary in order to account for the possibility of a split of an acoustic unit.

3.3 Feature Extraction:

The various properties that must exist in the features extracted from the speech signal for automatic speaker recognition are: high inter and low intra speaker variability, robust against error sources, easily measurable and maximally independent of each other.

The first property requires that feature extracted from a speech signal should be invariant with respect to the desired speakers while exhibiting a large deviation from the features of an impostor. In automatic recognition, the features must be measurable without the aid of a human expert. It has been found that a good feature is robust against several factors like voice disguise and distortion. The third property includes the requirement of two factors. Firstly, the feature should occur frequently and naturally in speech so that it could be extracted from short speech samples. Secondly, the feature extraction itself should be easy. Finally, different features extracted from the speech signal should be maximally independent of each other. If two correlated features are combined, nothing is gained, but this may degrade recognition results.

No feature has all the requirements listed above, but we can relax some of the requirements in automatic speaker recognition. In our task, we can forget the features that require human expert involved. In practice, the signal processing methods used in the feature extraction are computationally efficient. Some of the widely used features are predictor coefficients, cepstral coefficients and their derivatives, line spectral pairs (LSP), log area ratios (LAR), vocal tract area functions, and the impulse response of the filter. For speaker recognition, some of the above features were compared and the cepstral coefficients were found to provide the best results. Also, the derivatives of the cepstral coefficients capture the temporal information in speech that is essential for text dependent tasks. Here two commonly used feature extraction techniques are discussed for deriving cepstral features: one is LPC and other is MFCC.

3.3.1 Cepstral Coefficients:

The source filter model of speech production is shown in Figure 2.5. When air is forced through the vocal cords, then either the noise source is generated or the periodic pulses are generated, which are subsequently filtered by the shape of the vocal tract. Thus produced speech is split into a rapidly varying excitation signal and a slowly varying filter signal.

Let $s(t)$ denote the speech signal, $h(t)$ the impulse response of the LTI filter and $e(t)$ the excitation signal, then in the frequency domain,

$$S(f) = H(f) * E(f) \tag{3.4}$$

Since the envelope of the power spectra contains the vocal tract information, we deal $H(f)$ in a way to represent the envelope of the speech power spectra and $E(f)$ to represent the fine detail of the excitation. By taking logarithmic of both sides, a desirable separation of the excitation and vocal tract components is achieved, as

$$log\, S(f) = log\, H(f) + log\, E(f) \tag{3.5}$$

For most of speech processing, phase is not so important and we require only amplitude spectra [40], hence the above equation can be written as,

$$log\, (|S(f)|) = log\, (|H(f)|) + log\, (|E(f)|) \tag{3.6}$$

The slowly varying components of $log\, (|S(f)|)$ are represented by the low frequencies and the fine detail by high frequencies. Hence another Fourier transform is the natural way to separate the components of $H(f)$ and $E(f)$ as it preserves the sum. Thus most of the details occur near the origin and in peaks higher up the cepstrum. The overall shape of the log spectrum, i.e., the spectral envelope is described by the lower numbered cepstral coefficients and the remainder of the detail is contained in the higher coefficients. If f_0 is the frequency of the periodic pulse, the pitch period, then $log\, (|E(f)|)$ has peaks at integer multiples of f_0. These peaks translate into a relatively

steep bump in the cepstral domain and provide the fine detail pitch information. Thus produced cepstral analysis is shown diagrammatically in Figure 3.2.

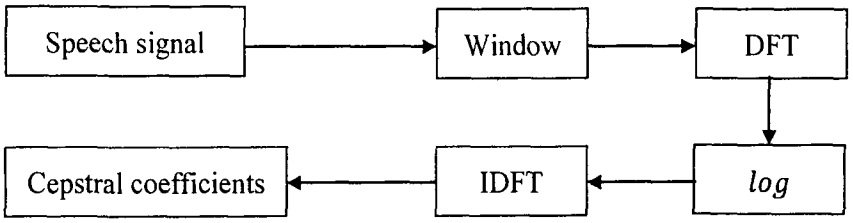


Figure 3.2: Block diagram presentation for cepstral coefficients extraction.

Cepstral analysis provides a method for separating out the vocal tract information from the excitation information. Thus the reverse transformation can be carried out to provide a smoothed power spectrum. This process is known as homomorphic filtering and can be used to remove linear time invariant channel effects. The cepstrum exponent due to such channel effects should be constant and can be subtracted out. Another benefit of using cepstra coefficients is that they can be reasonably modeled by multivariate Gaussian distributions [41]. Finally and perhaps most importantly, cepstra works well experimentally.

3.4 Linear Predictive Analysis:

Linear predictive analysis is historically one of the most important speech analysis techniques and also widely used in speaker recognition area. The basis of this technique is the source-filter model described in section 2.2.2, where it is assumed that this model is an all pole model. It was considered that the speech production system can be ideally characterized by the pole-zero system function and such assumption to use only poles has two main reasons. First reason is the simplicity, and as we will see that LPC will result in simple linear equations. Second reason is that based on human perception mechanism, human ear is fundamentally phase deaf and phase information is less important. All-pole model can exactly preserve magnitude spectral dynamics (the information) in the speech

but may not retain the phase characteristics. The way in which LPC is applied to the analysis of speech signals leads to a reasonable source-vocal tract separation. With sufficient parameters, this model can make a reasonable approximation to the vocal tract spectral envelope for all speech sounds.

This analysis performs a linear prediction of the next sample as the weighted sum of the past samples,

$$\tilde{S}_n = \sum_{k=1}^{p} a_k S(n-k) \tag{3.7}$$

Where the coefficients $a_1, a_2, \ldots\ldots\ldots\ldots, a_p$ are assumed constant over the speech analysis frame and are termed as predicton coefficients.

The transfer function of the linear filter is given as,

$$H(z) = \frac{1}{1-\sum_{k=1}^{p} a_k z^{-k}} \tag{3.8}$$

Parameter Estimation: One reasonable way to compute predictor coefficient a_k, is in terms of mean squared error. The prediction error (e_n) is defined as the difference between the real and predicted output, also called as prediction residual, at any time n is given as,

$$e_n = S_n - \tilde{S}_n \tag{3.9a}$$

$$= S_n - \sum_{k=1}^{p} a_k S(n-k) \tag{3.9b}$$

In speaker recognition task, we can use LPC based on the short term analysis approach. Because of the quasi stationary nature of speech, we can compute a set of prediction coefficient from every frame. Thus we can use these coefficients as features to describe the signal and therefore, the speaker. In practice, prediction order is set to 12-20 coefficients depending on the sampling rate. Thus the basic problem of linear prediction

analysis is to determine the set of predictor coefficients directly from the speech signal so that the spectral properties of the filter match those of the speech waveform within the analysis window. The basic approach is to find a set of predictor coefficients that minimize the mean squared error over a short segment of the speech waveform.

Hence, to determine the predictor coefficients, we define short term speech and error segments at time n as, $S_n(m) = S(n+m)$ (3.10a)

$$e_n(m) = e(n+m) \tag{3.10b}$$

Our next step is to minimize the mean squared error at time n a

$$E_n = \sum_m e_n^2(m)$$

$$= \sum_m \left[S_n(m) - \sum_{k=1}^{p} a_k S_n(m-k) \right]^2 \tag{3.11}$$

The minimization of E_n occurs when the derivative is zero with respect to each of the parameter a_k. Since the value of E_n is quadratic in each of the a_k therefore, there is a single solution. Very large positive or negative values of a_k must lead to poor prediction and hence the solution to $\frac{\partial E_n}{\partial a_k} = 0$ must be a minimum.

Hence, differentiating Eq. (3.11) with respect to a_k and setting equal to zero gives the set of p equations,

$$-\sum_m 2\left[S_n(m) - \sum_{k=1}^{p} \hat{a}_k S_n(m-k) \right] S_n(m-j) = 0 \tag{3.12}$$

$$-2\sum_m S_n(m)\, S_n(m-j) + 2\sum_m \sum_{k=1}^{p} \hat{a}_k S_n(m-k)\, S_n(m-j) = 0 \tag{3.13}$$

Rearranging the above equation gives,

$$\sum_m S_n(m)\, S_n(m-j) = \sum_{k=1}^{p} \hat{a}_k \sum_m S_n(m-k)\, S_n(m-j) \tag{3.14}$$

By recognizing that terms of the form $\sum S_n(m-j)S_n(m-k)$ are terms of the short term covariance of $S_n(m)$, i.e,

$$\varphi_n(i,k) = \sum_m S_n(m-i)\, S_n(m-k) \tag{3.15}$$

Hence we can write the Eq. (3.14) as,

$$\varphi_n(i,0) = \sum_{k=1}^{p} \hat{a}_k\, \varphi_n(i,k) \tag{3.16}$$

Hence the mean squared error E_n can be expressed as,

$$E_n = \sum_m S_n^2(m) - \sum_{k=1}^{p} \sum_m S_n(m)\, S_n(m-k) \tag{3.17a}$$

$$= \varphi_n(0,0) - \sum_{k=1}^{p} \hat{a}_k \varphi_n(0,k) \tag{3.17b}$$

Thus the minimum mean squared error consists of a fixed term $\varphi_n(0,0)$ and terms that depend on the predictor coefficients.

We have to compute $\varphi_n(i,k)$ for $1 \le i \le p$ and $0 \le k \le p$ by solving Eq. (3.16) for the optimum predictor coefficients and then solve the resulting set of p simultaneous equations. In general, the method of solving the equations is a strong function of the range of m used in defining both the section of speech for analysis and the region over which the mean squared error is computed.

The Autocorrelation Method:

One simple and straightforward way is to make the use of the fact that the samples are zero outside the interval $0 \leq m \leq N-1$. This is equivalent to assuming that the speech signal, $S(m+n)$, is multiplied by a finite length window, $w(m)$, which is identically zero outside the range, hence the speech segment can be represented as,

$$S_n(m) = \begin{cases} S(m+n).w(m) & 0 \leq m \leq N-1 \\ 0 & otherwise \end{cases} \tag{3.18}$$

Based on using the weighted signal, the mean squared error becomes

$$E_n = \sum_{m=0}^{N-1+p} e_n^2(m) \tag{3.19}$$

Hence, $\varphi_n(i,k)$ can be expressed as,

$$\varphi_n(i,k) = \sum_{m=0}^{N-1+p} S_n(m-i)\, S_n(m-k), \qquad 1 \leq i \leq p, 0 \leq k \leq p \tag{3.20a}$$

$$or, \quad \varphi_n(i,k) = \sum_{m=0}^{N-1-(i-k)} S_n(m) S_n(m+i-k), \qquad 1 \leq i \leq p, 0 \leq k \leq p \tag{3.20b}$$

Since $\varphi_n(i-k)$ is only dependent on the difference $(i-k)$ rather than two independent variables i *and* k, hence can be written in terms of the autocorrelation function as,

$$\varphi_n(i,k) = r_n(i-k) = \sum_{m=0}^{N-1-(i-k)} S_n(m)\, S_n(m+i-k) \tag{3.21}$$

Since the autocorrelation function is symmetric, i.e., $r_n(-k) = r_n(k)$ hence, the LPC Eq. (3.16) can be expressed as,

$$r_n(i) = \sum_{k=1}^{p} r_n|(i-k)|\hat{a}_k \tag{3.22}$$

or, in matrix form:

$$\begin{pmatrix} r_n(1) \\ r_n(2) \\ \vdots \\ r_n(p) \end{pmatrix} = \begin{pmatrix} r_n(0) & r_n(1) & \cdots\cdots & \cdots\cdots & r_n(p-1) \\ r_n(1) & r_n(0) & \cdots\cdots & \cdots\cdots & r_n(p-2) \\ \vdots & \vdots & & & \vdots \\ r_n(p-1) & r_n(p-2) & \cdots\cdots & \cdots\cdots & r_n(0) \end{pmatrix} \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_p \end{pmatrix} \quad (3.23)$$

The p × p matrix of autocorrelation values is a Toeplitz matrix (symmetric with all diagonal elements and equal) and hence can be solved efficiently through several well-known procedures. One of which is Durbin's algorithm.

LPC Processor:

The block diagram of LPC processor that has been widely used for recognition purpose includes the following steps, as shown in the Figure 3.3 given below:

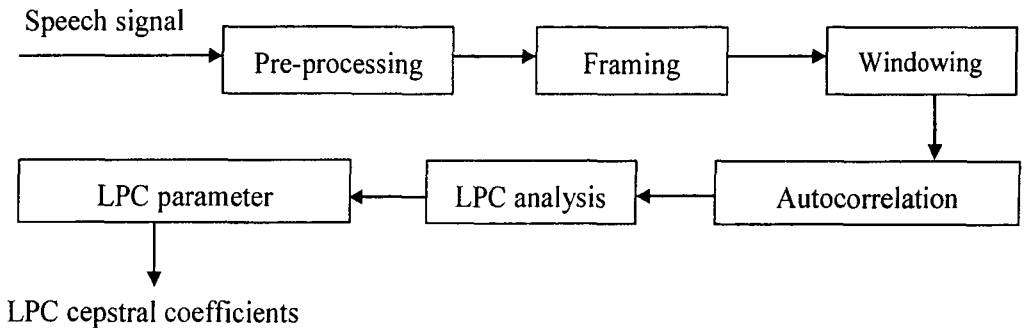


Figure 3.3: Block diagram of LPC processor [35].

Pre-processing: First step performs different steps of pre-processing of the measurement part that is common to all feature extraction techniques. In the next step, the signal is blocked into the frames of N samples with adjacent frames being separated by M samples such that, M ≤ N, and finally the windowing of each individual frame is done so as to minimize the signal discontinuities at the beginning and end of each frame.

Autocorrelation Analysis: Each frame of windowed signal is then autocorrelated to give,

$$r_l(m) = \sum_{n=0}^{N-1-m} x_l(n)\, x_l(n+m) \qquad ,m = 0,1,2, \ldots\ldots\ldots\ldots, p \tag{3.24}$$

where, the highest autocorrelation value, p, is the order of the LPC analysis. The benefit of the autocorrelation analysis is that the zeroth autocorrelation $r_l(0)$ is the energy of the l^{th} frame. The frame energy is an important concept for the speech detection system.

LPC Analysis: The next processing step is the LPC analysis that converts each frame of p+1 autocorrelations into a LPC parameter set. The formal method for converting from autocorrelation coefficients to a LPC parameter set is known as Durbin's method. Denoting the values of LP parameters at iteration i by $a_k^{(i)}$ and the residual energy by E_i for $i = 1,2, \ldots\ldots\ldots, p$.

$$E^{(0)} = r(0) \tag{3.25}$$

$$k_i = \left(r(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} r|(i-j)| \right) \Big/ E^{(i-1)} \tag{3.26}$$

$$a_i^{(i)} = k_i \tag{3.27}$$

$$a_j^{(i)} = a_j^{(i-1)} - k_i\, a_{i-j}^{(i-1)} \tag{3.28}$$

$$E^{(i)} = (1 - k_i^2)\, E^{(i-1)} \tag{3.29}$$

where, the coefficients $a_i's$ are LPC coefficients and $k_i's$ are known as the reflection coefficients.

LPC Parameter Conversion to Cepstral Coefficients: A very important LPC parameter set is used, because cepstrum is proved to be the most effective representation of speech signal for speaker recognition. The LPC cepstral coefficients C_m can directly be derived from the LPC coefficient set, using the following recursion:

$$C_1 = a_1 \tag{3.30}$$

$$C_m = a_m + \frac{1}{m}\sum_{i=1}^{m-1} iC_m\, a_{m-i} \qquad ,1 \le m \le p \tag{3.31}$$

$$C_m = \frac{1}{m}\sum_{i=1}^{m-1} iC_m\, a_{m-i} \qquad ,m > p \tag{3.32}$$

3.5 Non-Linear Scale Analysis:

The above analysis techniques place equal emphasis on every part of the frequency scale from zero up to the maximum representable frequency. There are two main reasons for using non-linear frequency scales. Firstly, it approximates the sensitivity of the human ear. Secondly, this analysis is used to get around the frequency/time resolution tradeoff. Using a narrow bandwidth at low frequency enables harmonics to be resolved but gives poor onset information and using a larger bandwidth at higher frequencies allows for high temporal resolutions of bursts, etc.

The pshychoacoustical studies have shown that the human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Taking auditory characteristics into account, the mel-frequency scale has been found the most commonly used scale from engineering point of view. This scale was projected by Stevens, Volkmann and Newman in 1937. It is divided into the units mel using the simple analytical relation as [42]:

$$m = 2595 * log10\,(1 + f/700) \tag{3.33}$$

where, f is the frequency in Hz and m is the resulting mel scaled frequency.

Here for each tone with an actual frequency f measured in Hz, a subjective pitch is measured on a scale called the 'Mel scale'. The Mel-frequency scale is a linear frequency spacing scale below 1000 Hz and logarithmic spacing scale above 1 KHz.

These filters are triangular in shape and overlapping in nature and are arranged linearly in the mel frequency domain. The filter arrangement in ordinary frequency domain is shown in the Figure 3.4 given below:

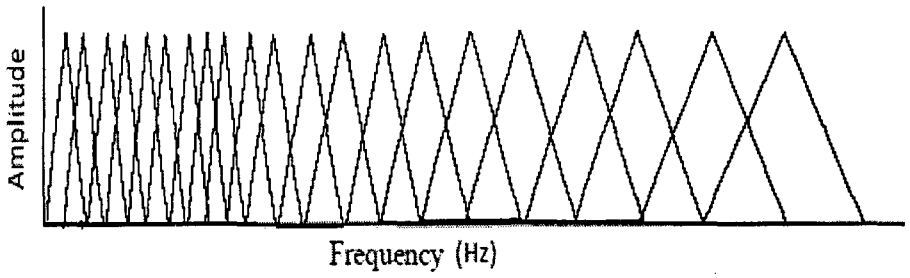


Figure 3.4: Mel-scale filterbanks [43].

3.5.1 Mel Frequency Cepstral Coefficients (MFCCs):

Mel cepstrum is one of the most commonly used feature extraction technique used in both speech and speaker recognition. This method reduces the frequency information of the speech signal into a small number of coefficients that emulate the separate critical bands in the basilar membrane of the ear, i.e., it tries to code the information in a similar way as human cochlea does. This technique is very popular in recognition as it has the basic desirable property that the coefficients are largely independent, allowing probability densities to be modeled with diagonal covariance matrices. Also, the Mel scaling has been shown to offer better discrimination between phonemes that is an obvious help in recognition. MFCCs extraction from the acoustical speech signal involves different steps as shown in Figure 3.5.

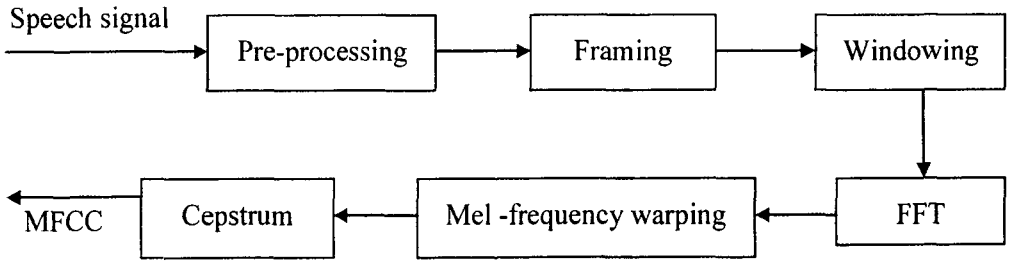


Figure 3.5: Block diagram of MFCC processor [44].

Pre-processing: First phase involves normalization, mean subtraction and pre-emphasis as we have discussed earlier in section 3.2. Further we perform framing of speech signal in order to analyze speech signal in shorter frames due to its non-stationary nature. The next step involves windowing of each frame that minimizes the discontinuity at start and end of each frame.

Fast Fourier Transform: This step involves the conversion of each windowed speech frame of N samples from time domain to frequency domain. The FFT is the fast algorithm to implement the DFT, which is defined on the set of N samples as:

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi nk/N} \qquad ,0 \leq k \leq N-1 \tag{3.34}$$

Mel Frequency warping: Passing the magnitude spectrum $X(k)$ through the mel-filter bank means that these magnitudes are then binned by correlating them with each triangular filter. Here binning means that each FFT magnitude coefficient is multiplied by the corresponding filter gain and the results accumulated. Thus, each bin holds a weighted sum representing the spectral magnitude in that filter bank.

Cepstrum: Since we want the coefficients to have the speaker specific 'vocal tract' characteristics in them, hence we try the cepstral coefficients that provide a good representation of the local spectral properties of the signal for the selected frame analysis.

The cepstral characteristics in features is included by taking the logarithmic of the filter bank outputs followed by Discrete Cosine Transform (DCT) to convert spectrum back to time domain.

Because the mel spectrum coefficients (and their logarithm) are real numbers, we can convert them to the time domain using the DCT. The advantage of DCT is that it de-correlates the features and arranges them in descending order of information. Therefore, the MFCCs are calculated from the logarithm of mel power spectrum coefficients of last step, denoted $m_j, j = 1,2, \ldots\ldots\ldots, N$, using DCT as [42]:

$$C_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{\pi i}{N}(j - 0.5)\right) \tag{3.35}$$

where, $i = 1, \ldots\ldots\ldots\ldots$, number of cepstral coefficients and N is the number of filterbanks.

This spectrum provides a fairly simple but unique representation of the spectral properties of the speech signal. Thus, we get the MFCCs for each speech frame by applying the procedure described above. This set of coefficients is called the acoustic features and hence each input utterance is transformed into a sequence of acoustic features. In the next section, we will discuss that how these features can be used to represent and recognize the voice characteristics of the speaker.

3.6 Classifier:

Three stages are generally involved in building a pattern recognition system: training, testing and decision. In the training stage, a set of parameters of the model is estimated so that in some sense the model learns the correspondence between the features and the levels of the objects. In the testing stage, the parameters of the model are then adjusted to achieve a good generalization of the performance of the system. This stage usually consists of a set of features and levels that are different from the training data. The task of

recognition is carried out in the decision stage where, the features with an unknown label are passed through the system and assign a level at the output.

The statistical pattern recognition can be summarized in a single word as 'classification'; supervised (or discrimination) and unsupervised (simply referred to as classification or clustering) classification. In supervised classification, we have a set of data samples (each consisting of measurements on a set of variables) with associated labels, the class types. In unsupervised classification, the data are not labeled and we try to find groups in the data and features that distinguish one group from another. Based on the classification criterion used in the discriminant functions, classifier can be grouped into Bayesian classifier, Likelihood classifier and distance classifier [45]. This technique has a number of applications in various fields as from automatic character recognition and medical diagnosis to the development of machines with brain like performance that in some way would emulate human performance.

3.6.1 Discriminant Analysis:

Discrimination comes into the category of supervised classification. In discrimination, we assume that there exists C classes or groups denoted by $w_1, w_2, \ldots\ldots\ldots\ldots, w_C$ and associated with each pattern x is a categorical variable z that denotes the class or group membership, that is if, $z = i$, then the pattern belongs to w_i, $i \epsilon (1,2, \ldots\ldots\ldots, C)$. Elements of patterns are measurements of an acoustic waveform in speech and speaker recognition problems. For discrimination purpose, we assume that we have a set of patterns of known class $\{(x_i, z_i);\ i = 1, 2, \ldots\ldots\ldots\ldots, C\}$ (the training set) that we use to design the classifier. Once this has been done, we may estimate class membership for an unknown vector x. The patterns are sequences of acoustic features in our case that we have extracted from an input speech using the technique described in the previous section. The classes here refer to individual speakers. Since the classification process in our case is applied on extracted features, it can also be referred to as feature matching.

3.6.2 Linear Discriminant Analysis (LDA):

The LDA method consists of searching some linear combinations of selected variables, which provide the best separation between the considered classes. These different combinations are called discriminant functions. Discriminant analysis method was originally developed in 1936 by R. A. Fisher. This method often produces models whose efficiency approaches (and occasionally exceeds) more complex modern methods. LDA is used to perform dimensionality reduction while preserving as much of the class discriminatory information as possible. It tries to find the direction along which the classes are best separated by taking into consideration the scatter within class but also the scatter between classes.

Let there are C classes in the training data. If μ_i be the mean vector of the class i and m_i be the number of samples within class i, where, $i = 1,2, \ldots \ldots, C$, then

Total number of samples will be,

$$M = \sum_{i=1}^{C} m_i \tag{3.36}$$

And, mean vector of the entire data set will be,

$$\mu = \frac{1}{C}\sum_{i=1}^{C} \mu_i \tag{3.37}$$

Now within class scatter matrix S_w and between classes scatter matrix S_b can be defined as,

$$S_w = \sum_{i=1}^{C}\sum_{j=1}^{m_i} (y_j - \mu_i)\,(y_j - \mu_i)^T \tag{3.38}$$

$$S_b = \sum_{i=1}^{C} (\mu_i - \mu)\,(\mu_i - \mu)^T \tag{3.39}$$

LDA computes a transformation that maximizes the between class scatter while minimizes the within class scatter in such a way, maximize $\frac{\det(S_b)}{\det(S_w)}$. This transformation tries to rotate the axes so that when the categories are projected on the new axes, the differences between the groups are maximized as illustrated by the Figure 3.6.

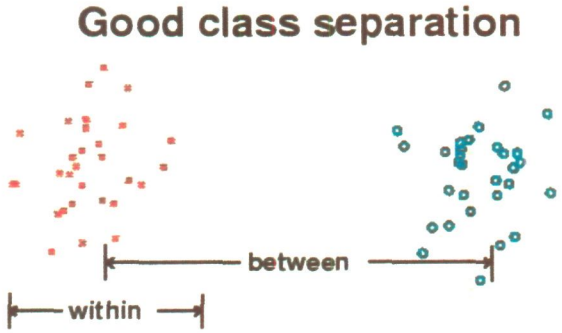


Figure 3.6: Example of LDA classifier produced by Ludwig Schwardt and Johan du Preez [46].

3.6.3 Principal Component Analysis (PCA):

Data is often described by many more variables then necessary for building the best model. Sometimes, specific techniques exist for selecting a good subset of variables, but dimensionality reduction technique such as PCA may also be considered for feeding the model with a reduced number of variables. PCA describes smaller set of variables that explain most of the variance in the original data, in more compact and insightful form. PCA was invented in 1901 by Karl Pearson. Now it is mostly used as a tool in exploratory data analysis and for making predictive models.

This technique is the form of unsupervised learning. Geometrically, it can be viewed as a rotation of the existing axes to new positions in the space defined by original variables that have the following properties:

- Ordered such that the principal axis 1 has the highest variance, axis 2 has the next highest, ..., and axis p has the lowest variance.
- Covariance among each pair of the **p** axes is zero (the principal components are uncorrelated).

Quite generally, reducing the number of variables used to describe data will lead to some loss of information. PCA operates in a way that makes this loss minimal, in a sense that will be given a precise meaning. The properties of the principal components are:

Number: Although the ultimate goal is to use only a small number of principal components. PCA first identifies **p** such components, that is, the same number as the number of original variables. Only later will the analyst decide on the number of components to be retained. "Retaining **p** principal components" means "Replacing the observations by their orthogonal projections in the **p**-dimensional subspace spanned by the first **p** principal components".

For **n** original dimensions, correlation matrix is **n** × **n** and has **n** eigen vectors, so **n** principal components. But if the eigen values are small, we can ignore the components of lesser significance and then we can choose first **p** eigen vectors based on their eigen values. Hence, final data has only **p** dimensions.

Orthogonality of the principal components: The principal components define orthogonal directions in the space of observations. In other words, PCA just makes a change of orthogonal reference frame, the new variables being replaced by the principal components.

Uncorrelatedness of the principal components: It will turn out that the principal components are pariwise uncorrelated.

Ordering of the principal components; optimal projection subspaces: The fundamental property of the principal components is that they can be ordered by decreasing order of importance.

3.7 Neural Network Classifier:

Two key concepts of artificial intelligence are automatic knowledge acquisition (learning) and adaptation. One way in which these concepts have been implemented is via the neural network approach [35]. A neuron is an information processing unit that is fundamental to the operation of a neural network. The block diagram given in Figure 3.7 shows the model of a neuron, which forms the basis for describing neural networks. The basic three elements of a neuronal network are defined as follows:

The first element is a set of synapses or connecting links, each of which is characterized by a weight or strength of its own. Specifically, signal x_j at the input of synapse j connected to neuron k is multiplied by the synaptic weight w_{kj}, where the first subscript refers to the neuron and the second subscript refers to the input end of the synapse to which the weight refers. Unlike a synapse in the brain, the synaptic weight of artificial neuron may lie in a range that includes negative as well as positive values.

The second element is an adder for summing the input signals, weighted by the respected synapses of the neuron (linear combiner).

The last element is an activation function for limiting the amplitude of the output of a neuron. The activation function is also referred to as a squashing function in that it squashes (limits) the permissible amplitude range of the output signal to some finite value.

The natural model also includes an externally applied bias b_k, denoted by. The bias b_k has the effect of increasing or lowering the net input of the activation function, depending on whether it is positive or negative respectively.

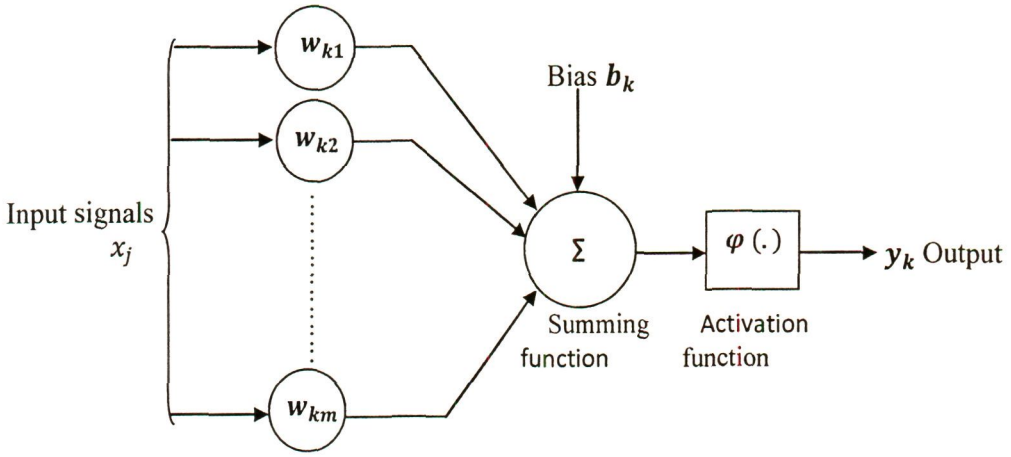


Figure 3.7: Computational elements of a Neural Network [47].

In mathematical terms, we describe a neuron k by writing the following pair of equations:

$$u_k = \sum w_{kj}\, x_j \qquad j = 1, 2, \ldots\ldots\ldots\ldots, m$$

$$y_k = \varphi\,(u_k + b_k)$$

where, x_1, x_2, ..., x_m are input signals, w_{k1}, w_{k2}, ..., w_{km} are the synaptic weights of neuron k, u_k is the linear combiner output due to the input signals; b_k is the bias, $\varphi(.)$ is the activation function and y_k is the output signal of the neuron.

Neural Network:

A neural network is also called a connectionist model, a neural net or a Parallel Distributed Processing (PDP) model. It is made up of simple processing units, which has a natural tendency for storing experimental knowledge and making it available for use. It resembles the brain in two respects: Knowledge is acquired by the network from its environment through a learning process and interneuron connection strengths, known as synaptic weights, are used to store the acquired knowledge. The biological basis of the neural network is a model by McCullough and Pitts of neurons in the human nervous system and it exhibits all the properties of neural element, as shown in Figure 3.7.

Feedforward Neural Network:

A feedforward neural network is a biologically inspired classification algorithm. It consists of a large number of simple neuron like processing units, organized in layers. Every unit in a layer is connected with all the units in the previous layer. These connections are not all equal and may have a different strength or weight. The weights on these connections encode the knowledge of a network. Often the units in a neural network are also called nodes. Data enters at the inputs and passes through the network, layer by layer, until it arrives at the output. During normal operation that is, when it acts as a classifier, there is no feedback between layers. This is why these are called feedforward neural networks.

Backpropagation Neural Network:

The differences between the actual outputs and the idealized outputs are propagated back from the top layers to lower layers to be used at these layers to modify the connection weights.

The Figure 3.8 shows a 3- layered network with, from top to bottom: an output layer with 4 units, two hidden layers each with 5 and 6 units respectively. The network has 3 input units. The 3 input units are shown as circles and do not belong to any layer of the network. Any layer that is not an output layer is a hidden layer. This network therefore, has 2 hidden layers and 1 output layer. The figure also shows all the connections between the units in different layers. A layer only connects to the previous layer.

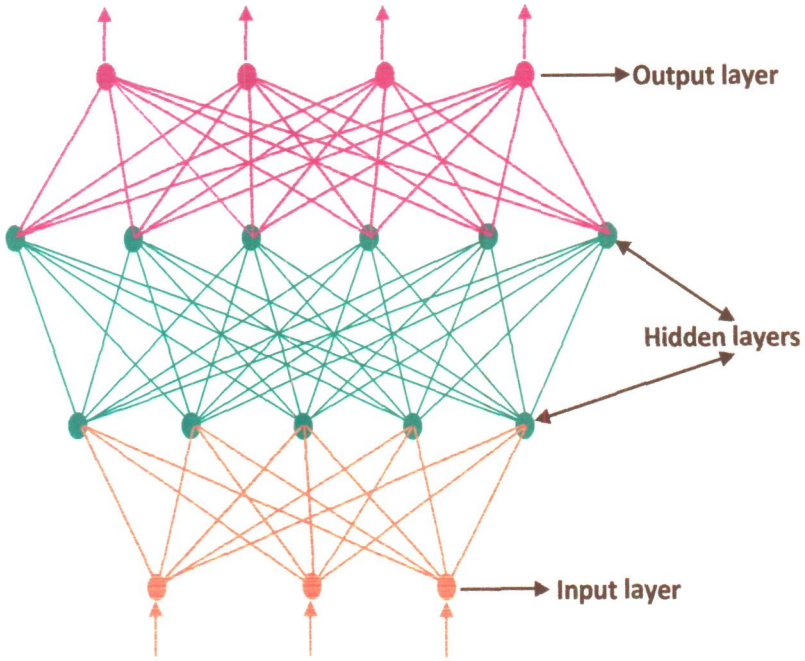


Figure 3.9: Three layered Neural Network

The operation of this network can be divided into two phases:

The Learning Phase: The process used to perform the learning process is called a learning phase, the function of which is to modify the synaptic weights of the network in such a way that when a pattern is presented, the output unit with the correct category, hopefully will have the largest output value.

In supervised learning, the learning process takes place under the guidance of a teacher. However, in the paradigm known as learning without a teacher, as the name implies, there is no teacher to oversee the learning process. That is to say, there are no labeled examples of the function to be learned by the network.

Learning by Feedforward Neural Network:

The FFNet uses a supervised learning algorithm: besides the input pattern, the neural net also needs to know to what category the pattern belongs. Learning proceeds as follows: a pattern is presented at the inputs. The pattern will be transformed to the output layer through the layers of the network. The units in the output layers all belong to a different category. The outputs of the network as they now compared with the outputs as they ideally would have been if this pattern were correctly classified: in the later case the unit with the correct category would have the largest output value and the output values of the other output units would have been very small. On the basis of this comparison all the connection weights are modified a little bit to guarantee that, the next time this same pattern is presented at the inputs, the value of the output units that corresponds with the correct category is a little bit higher than it is now and that, at the same time, the output values of all the other incorrect outputs are a little bit lower than they are now. If the above procedure is performed once for every pattern and category pair in the data set, then it means that 1 epoch of learning is performed.

Maximum number of epochs means the maximum number of times that the complete pattern data set will be presented to the neural net. The hope is that eventually, probably after many epochs, the neural net will remember these pattern–category pairs. We even hope that the neural net when the leaning phase has terminated, will be able to generalize and has learned to classify correctly any unknown pattern presented to it. Because real life data many times contains noise as well as partly contradictory information these hopes can only be partly fulfilled.

Learning Curves: The learning curve is a plot of the mean square value of the estimation error, i.e., the difference between the desired response and actual filter output, called cost function, versus the number of iterations.

Classification Phase: In the classification phase the weights of the network are fixed. A pattern, presented at the inputs, will be transformed to the output layer from layer to layer

until. Now classification can occur by selecting the category associated with the output unit that has the largest output value.

Advantages of Neural Network:

The neural networks are widely used for a range of problems for several reasons as:

i. A neural network is a highly parallel structure of simple, identical, computational elements, hence can readily implement a massive degree of parallel computations.

ii. Since the information embedded in the neural network is spread to every computational element within the network, this structure is inherently among the least sensitive of networks to noise or defects within the structure.

iii. The concept of adaptive learning is inherent in the neural network structure which means that the connection weights of the network need not be constrained to be fixed; they can be adapted in real time to improve performance.

iv. Neural networks provide inputs and outputs because of the non-linearity within each computational element and hence are often more efficient than alternative physical implementations of the nonlinearity.

CHAPTER 4

IMPLEMENTATION AND RESULTS

4.1 Database Preparation:

Preparing a database is the most crucial task in the recognition system. The database management is concerned with the recording of the audio files. The speaker recognition task in this dissertation is tested on English and Hindi language speech databases.

4.1.1 English Database:

In 2004, English Language Speech Database called ELSDSR was prepared especially to Speaker Recognition applications [48]. The main characteristics of this database are: English spoken by non-native speakers, a single-session of sentence reading and relatively extensive speech samples suitable for learning person specific speech characteristics. ELSDSR contains voice messages from 23 speakers (13 Male and 10 Female), age from 24 to 63 years at Technical University of Denmark. In this database, 11 sentences arranged in 7 paragraphs were read by each speaker for training set (given in List (I) of Appendix-A). These sentences were made with an attempt to capture all the phonemes of English language (vowels, consonants and dipthongs, etc.). Forty six sentences (different from the training set sentences given in List (II) of Appendix-A) were recorded for the test set and 2 sentences were read by each (23) speaker.

Using this database, 48-dimensional MFCC were taken as the desired features and speaker pruning technique was introduced into the recognition system for the purpose of increasing the recognition rate [49]. K-nearest neighbor (KNN) classification method and Discrete Density Hidden Markov Model (DDHMM) were used for recognition. The

highest recognition performance of the designed speaker recognition system was achieved 92.07%.

This database was further used in 2010 [50]. A novel text-independent speaker identification system was implemented using Zak transform coefficients as a feature set and resulted 100% recognition rate in case of using both the two sentences of the test set in identifying all the 23 speakers of this database. In this work, the Zak method was compared to MFCC method and a clear advantage was found in both modeling and identification complexity.

4.1.2 Hindi Database:

It would be more practical if one could collect the speech data from daily natural conversation. But we have used 10 (0 – 9) digits **(In IPA - / ʃunj /, / ek /, / do /, / tin /, / tʃar /, / pantʃ /, / tʃʰɛ /, / sat /, / aʈʰ /, / nɔ /)** in Hindi language, as it is frequently used in the applications of speaker recognition. The digits were spoken by 21 speakers (15 male and 6 female) and were recorded using "Cool Edit" software at 16 KHz sampling rate. Each digit was repeated 10 times by all the 21 speakers. Finally, database consists of $21 \times 10 \times 10 = 2100$ samples of Hindi digits.

4.2 Implementation:

As discussed in the previous chapter, the task of speaker recognition can be performed using the following three steps: pre-processing, feature extraction and classification.

4.2.1 Pre-processing:

After performing pre-emphasis (pre-emphasis constant equal to 0.97), mean subtraction and normalization steps, the speech sample is blocked into samples. We have selected the number of samples equal to 256 corresponding to 16 ms frame size at 16 KHz sampling rate. We have also chosen an overlapping of 50%. Thus the second frame starts after the

8ms of first frame. In the next step, each frame is multiplied by equal size of Hamming window, and then feature extraction process is performed on each frame.

4.2.2 Feature Extraction:

Feature extraction technique employed is MFCC using "MATLAB" software. This technique results the well known features used to describe the speech signal, which are based on the known evidence that the information carried out by low frequency components of the speech signal is phonetically more important for human than carried by high frequency components. The technique of computing MFCC is based on short term analysis and thus from each frame a MFCC row vector is computed.

Since every utterance gives more than one row of MFCC vectors, depending on its length because all the sentences are used with their different lengths without any normalization. These coefficients are arranged successively in a matrix of size L × C, where L represents the number of rows equal to the number of frames and C represents the number of columns equal to the number of MFCCs extracted. Then the mean of the column matrix is computed and finally 1 × C matrix is obtained representing the model for that speaker.

For MFCC extraction, two important points are needed to be discussed: the number of bandpass filters in Mel scale, and the dimension of MFCCs extracted. Twenty four or 30 bandpass filters with 13-dimensional MFCCs have been commonly used for speech recognition [51, 39]. The case of speaker recognition is different from speech recognition, as the speaker information is not uniformly encoded in frequency bands. As the information of glottis is mainly encoded in a low frequency band (between 100Hz to 400Hz) and the information of the side branch (of articulators) is in high frequency band (between 4KHz to 5 KHz) [29]. The important characteristic of these frequency regions is that they show large variations among speakers, but have small changes during speech production for the same speaker. Also, these side branches are not easily disguised in the speech. Furthermore, the most speech phonemic discriminative information, such as the

first 3 formants is encoded in low and middle frequency regions (200 Hz to 3 KHz). In the paper [52], 60 bandpass filters were used to integrate each frequency band to get the Mel power spectra. Previous studies on speaker recognition have obtained the good recognition rate using 38- and 48- dimensional MFCC features [28, 49]. Further, the number of features was extended to 60 to make an improvement in the recognition rate [52].

4.2.3 Classifier:

The speaker identification experiment is conducted on ELSDSR and Hindi digit databases using LDA classifier. In case of ELSDSR database consisting of 23 speakers, for training set 161 (7 × 23) and for test set 46 (2 × 23) speech sentences are used. In case of Hindi digit database, 8 digits (0 - 7) with its 10 repetitions by each speaker, i.e., 21 × 8 × 10 = 1680 digit samples are selected for training set. Rest 2 digits (8 and 9 digits) with its 10 repetitions by each speaker, i.e., 21 × 2 × 10= 420 samples are selected for test set. The "MATLAB" program is used to evaluate the average percentage recognition (Appendix-B).

4.3 Results:

4.3.1 In Case of ELSDSR Database:

The average percentage recognition is obtained using LDA classifier for 24, 30 and 60 Mel filterbanks in the frequency range 0-8 KHz with different dimensions of mean value of MFCC features, as given in Table 4.1. The Mel-spaced filterbank outputs are converted into MFCCs by taking log of the filterbank outputs followed by DCT. The first coefficient is excluded since it carries a little speaker specific information.

In case of 24 Mel filterbanks, the recognition rate is calculated using 13, 18 and 23 MFCC features (from row 1-3 in Table 4.1). Further, in case of 30 Mel-spaced filterbanks, the average percentage recognition is obtained using 13, 18, 24 and 29 MFCC features (from row 4-7 in Table 4.1). For 60 Mel-spaced filterbanks, 13, 18, 24, 38, 48 and 52 dimensional MFCCs are taken for calculating average recognition score. Figure 4.1(a) represents the graphical representation for the variation in recognition rates with different Mel-spaced filterbanks and different MFCC features. It is clearly observed that the better average percentage recognition results using 60 Mel-spaced filterbanks.

Table 4.1: Average Percentage Recognition Results for ELSDSR database.

S.No.	Number of band pass filters in Mel-scale	Dimension of mean value of MFCC as feature vectors	Average Percentage Recognition
1	24	13	91.30
2	24	18	90.24
3	24	23	91.30
4	30	13	93.48
5	30	18	93.48
6	30	24	89.13
7	30	29	91.30
8	60	13	100
9	60	18	100
10	60	24	100
11	60	38	95.65
12	60	48	97.48
13	60	52	100
14	60	59	100

4.3.2 In Case of Hindi Digit Database:

The same feature extraction technique with LDA classifier is applied on Hindi digit database. Since we have found the good recognition score for ELSDSR database using 60 Mel-spaced filterbanks, due to this reason we have calculated the recognition performance for Hindi digit database with only 60 Mel-spaced filterbanks. The recognition rate is obtained for different dimensions of MFCCs and the good performance is achieved for higher dimensional (52-dimensional) MFCCs. Hence, in the next section onwards only the recognition performance for all the speakers with 52-dimensional MFCCs will be discussed. The variation with increasing dimension of MFCCs using 60 Mel-spaced filterbaks is shown in Figure 4.1(b).

Table 4.2: Percentage Recognition Results for Hindi digit database with LDA classifier

Number of Mel-spaced filterbanks	Dimension of MFCC features	Average Percentage Recognition
60	13	64.05
60	18	70.17
60	24	73.09
60	38	75.19
60	48	79.05
60	**52**	**80.24**
60	59	76.90

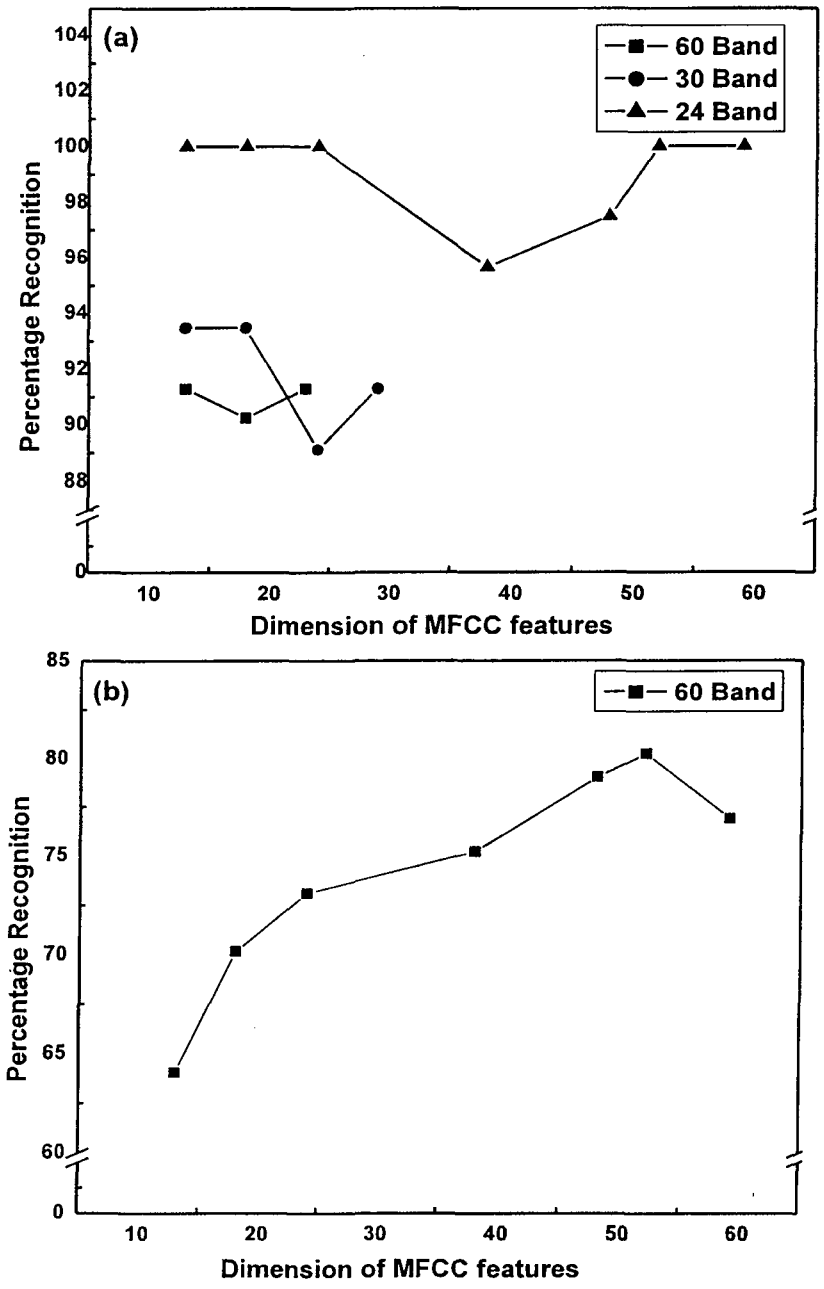


Figure 4.1: (a): Graphical representation of percentage recognition for ELSDSR database using different number of filterbanks and different dimensions of MFCCs.
(b): Graphical representation of percentage recognition for Hindi digit database using 60 Mel-spaced filterbanks and different dimensions of MFCC features.

4.4 Confusion Matrix Analysis:

The classification analysis produces a n x n confusion matrix for n-class problem that compares test versus train group membership. This matrix is the measure of how well the discriminant functions recognize group membership. But if a classification analysis with unknown grouping of objects is considered, then a confusion matrix can not be constructed. This confusion matrix shows the correct and incorrect classification. An analysis of the aggregate confusion matrix represents the overall classification and the analysis of the individual rows of the confusion matrix suggests that how significantly each respective group is classified than other expected groups and which groups are best classified by the discriminant function. In this matrix, individual cell represents the significant difference between the proportion of subjects correctly classified or misclassified within each group.

Confusion matrices for test obtained from 52-dimensional MFCC with 60 Mel-spaced filters using discriminant classifier for ELSDSR and Hindi digit databases are given in Table 4.3 and 4.4 respectively. The average result of recognition of speakers in case of ELSDSR database is 100% and of digit database is 80.24%.

4.5 Conclusion:

It is clear from the confusion matrix of ELSDSR database that using the 2 sentences of the test data for the recognition of speaker, it is found that all the 23 speakers are correctly identified by 100% (both the 2 sentences). But in Hindi digit database, 20 samples of the test data are used for the identification of speakers. In this case, some speakers I, J and O are identified correctly by 100% (20 samples are correctly classified) while speakers S and P are identified poorly by 45% and 30% (i.e., 9 and 6 samples are classified correctly) respectively. It suggests that sometimes a speaker is highly or less confused with other speakers, such as speaker P is more confused with speakers N and C

by 25% while less confused with speakers H and S by 5%, and thus the reduction is found in the recognition rate.

In this study, we first analyzed the percentage recognition results for various observations using MFCC features on ELSDSR database and then applied the best observed results with the same feature extraction technique on digit database. But the recognition rate that is observed for digit database is not as good as for ELSDSR database.

CHAPTER-5

FURTHER STUDIES USING HINDI DIGIT DATABASE

To improve the recognition performance, we have two choices, either including any other feature for classification or improving the classification technique.

5.1 Inclusion of Other Features with MFCC Features:

Since the nasal or nasalized characteristic is also an important speaker information and is used for speaker recognition by human beings, but it was observed in [30] that the statistical analysis could not find the correlation of nasal characteristics with speaker identity. The LPC model is an all pole model that can capture the resonant frequencies or formants but not the zeros, which are important for the nasalized sounds. But a serious problem with LPC is that they are highly correlated, however, it is desirable to obtain less correlated features. Hence, LPC cepstral coefficients are used to decorrelate the LP coefficients. Both the MFCC and LPC are well known techniques used in speaker identification to describe signal characteristics relative to the speaker discriminative vocal tract properties. They are quite similar as well as different and results in cepstrum coefficients, but the method of computation differs. There is no general agreement in the literature about what method is better. However, it is generally considered that LPCCs are computationally less expansive while MFCCs give precise result.

The main point of consideration in LPC analysis is the prediction order. We have chosen the value of prediction order according to the thumb rule, i.e., 2 + (fs/1000) = 18, and then 18 LPCC coefficients are used as feaures. But as we have said earlier that, the

first cepstral coefficient carries little speaker specific information, which may be responsible for the degradation of the recognition score. Hence, finally 17 LPCCs are chosen for the classification purpose. Again we compute the mean value of LPCC coefficients and finally get 1 x C matrix, where C corresponds to the number of columns (the number of LPCC coefficients extracted, as in our case it is 17).

Now our next step is to improve the classification technique. Here multi-layer Neural network and LDA with combination of PCA are used for classification.

5.2 Effect on Recognition Rate using LPCC and MFCC Features with LDA Classifier:

The average recognition percentage is improved by 1.9% using both MFCC and LPCC features for classification as compared to the recognition percentage obtained using MFCC features alone with LDA classifier. It is found from the confusion matrix obtained in this case (given in Table 5.1 of Appendix-B) that only for the speakers A, F and U, the correctly classified digit samples of the test set (20) are less than correctly classified test digit samples obtained from using MFCC features alone. While for all other speakers, either the correct classification of test samples is same or better.

In our next step, two other methods Neural network classifier and PCA dimensionality reduction technique is used for better performance of the recognizer system.

5.3 Neural Network Classifier:

By using measured features (as in our case MFCCs and LPCCs), we have attempted recognition of speakers with Neural network classifier using "PRAAT" software. To train and test our algorithms, we first split our data into training and test subset. Then we create a "TableOfReal" for both training and test set. A TableOfReal object contains a number of cells. Each cell belongs to a row and a column. For instance, a TableOfReal with 10 rows and 3 columns has 30 cells. Each row and each column may be labeled with

a title. Now we create pattern and categories for both (training and test set) TableOfReal. This is done by selecting table of real and then the option "To Pattern and Categories" from the dynamic menu, that is available in the "Convert" action button. Here Pattern refers to the input value, i.e., the values of different features for different categories and "Categories" refers to the output value (in our case, there are 21 categories representing speakers).

Now select pattern and categories together for trainng set, and select "To FFNet". A 3- layered feedforward neural network (2 hidden layers and 1 output layer), with 40 – 40 nodes in the hidden layers and 21 nodes in the output layer, is used. The number of nodes in the output layer corresponds to the number of categories. First, the result of the classifier is observed using only 52-dimensional MFCC features, thus we have 52-element input vector as the number of elements in the input vector corresponds to the number of features extracted. In this way, a FFNet as a classifier is built. For learning, we have to select 3 different objects together: a FFNet (the classifier), a Pattern (the inputs) and a Categories (the outputs) and then choose learn option. By doing this, the FFNet classifier will learn (trained) the categories. For learning (training), we have to choose number of epochs (i.e., the number of iterations for which the complete data set will be represented to network). We have selected epochs as 8000 with 4 times repetition. The expectations is that eventually, probably after many epochs, the neural network will remember pattern-category pairs. We have selcted these values of nodes for hidden layers and epochs because below these values the network was not able to learn the categories and above these values there was not much increment in performance.

For classification of speakers from the test data, we first create pattern and categories from the test set. Now select the FFNet (which has been trained) and pattern together, and then select "To Categories". Now a different category object appears. Now we select this category object and he previous category object (of the test set) together and choose "To Confusion" object and the value of fraction correct is obtained from

"Confusion: Get fraction correct". We have also drawn confusion table by selecting "Confusion" object and then selecting "draw", as given in Table 5.2 (Appendix-B).

Futher, the classification of speakers is evaluated using both 52-dimensional MFCC and 17-dimensional LPCC features. In this case, only the input units are changed, as now we have 69 input units. The conditions (hidden nodes and epochs) are unchanged, as these give the better results for this case also. The confusion matrix thus obtained from the test data is shown in Table 5.3 (Appendix-B).

5.3.1 Results and Disussion:

The average percentage recognition of speakers is 74.76% in case of operating only MFCC features extracted from the speech signal and 80.24%, when operating both the MFCCs and LPCCs extracted. It is observed from the results obtained using neural network classifier that although the average percentage recognition is 5.48% improved using both the LPCCs and MFCCs features for classification, but the recognition rate is not improved either using only MFCC features or both the MFCC and LPCC features than obtained with LDA classifier.

Though the capabilty of neural networks to discriminate between patterns of different classes is exploited for speaker recognition, but the disadvantage of this method of classification is that it takes more time for training the network. In our case, the neural network leads no improvement in the performance of the recognizer. It might be possible this classifier could give the better recognition results if we would use this method for text dependent speaker identification, as it was reported in [20] that neural network provides efficient results for dependent task.

5.4 Classification using LDA with PCA:

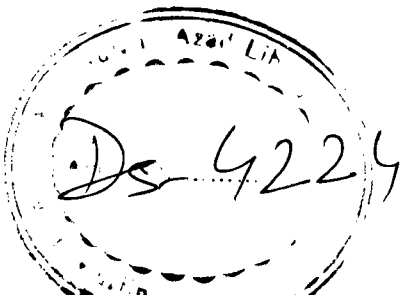
PCA constructs a low dimensional representation of the data that describes as much of the variance in the data as possible. This dimensionality reduction technique is applied on

the data with the help of "MATLAB" software using the program (given in Appendix-C). When LDA with PCA is applied on test data using only MFCC features extracted from the speech signal, it uses only 37 principal components out of 52 features and improvement of 0.95% is found in the recognition percentage as compared to using singly LDA classifier. But the recognition is found 6.43% better from using Neural network classifier. The reason of choosing 37 principal components is that it gives better performance as compared to selecting less or more number of principal components. The confusion matrix obtained in this case is given in Table 5.4 (Appendix-B). Further, when MFCC and LPCC features are used as the inputs for first dimensionality reduction using PCA and then classification using LDA, means 52 principal components are used out of 69 features, then the reconition percentage is improved by 1.67%. The confusion matrix thus obtained for the test set is given in Table 5.5 (Appendix-B).

Here it would be better to state that this technique gives further improvement over both LDA and Neural network classification techniques. Dimensionality reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality. Due to these reasons, LDA with combination of PCA is more adequate for speaker recognition than other singly used classification techniques like LDA and Neural network.

5.5 Different Approaches Applied on Hindi Digit Database for the improvement of the Speaker Recognition System:

Firstly, the recognition score was calculated using MFCC features with LDA classifier. Further, to improve the recognition rate some modifications are done by using MFCC with LPCC features and other classifiaction techniques as Neural network and LDA with PCA. Thus a comparative study using different classification techniques with using once MFCC alone and then using MFCC and LPCC both is done for calculating the better



average pecentage recognition of speakers using Hindi digit database, as given in Table 5.6 (shown in Figure 5.1).

Table 4.12: Comparison in recognition performance using Hindi digit database.

Features Extracted	Classification Technique	Average Recognition (%)
MFCC	LDA	80.24
MFCC + LPCC	LDA	82.14
MFCC	NNet	74.76
MFCC + LPCC	NNet	80.24
MFCC	PCA (37) + LDA	81.19
MFCC + LPCC	PCA (52) + LDA	**83.81**

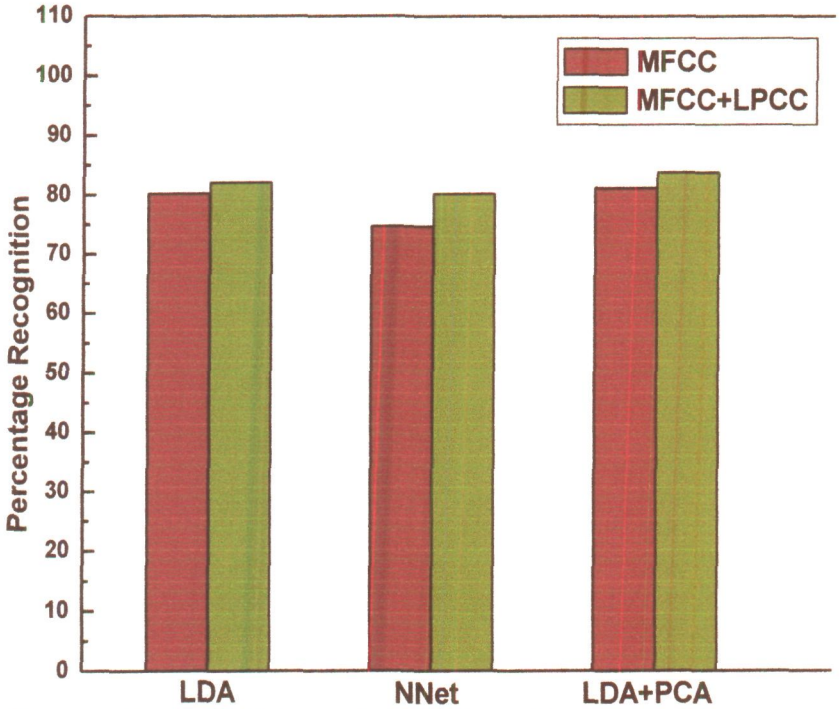


Figure 5.1: Comparative analysis for evaluating the performance of the recognizer on Hindi digit database using different features with various classification techniques

CONCLUSION

This dissertation addresses mainly two issues involved in the recognition of speakers in text independent case. These issues are to explore the dimensions of MFCC features and choose the better classifiaction technique that improves the recognition performance as well as reduces the computational cost. It has already been suggested from the speech production point of view that speaker information is deeply concerned with physiological and morphological differences of the speech organs and glottis causes dominant speaker discriminative information in low frequency region. The recognizer is designed for the text-independent environment by using the different text for both training and testing set, i.e., the text data in the test set is not used for the training. Modern speaker recognition applications require high accuracy at low complexity and easy calculations. Firstly, the observations using MFCC features are applied on ELSDSR database using LDA classifier so that the results are made comparable. Now the same structure of the recognizer is applied on Hindi digit database. Further, to improve the recognition rate some modifications are done by using MFCC with LPCC features and other classification techniques like Neural network and LDA with PCA.

Finally, from all the results we conclude the following factors:

First, the modification in the MFCC technique (mean vlaue of MFCC) using 60 Mel-spaced filterbanks and LDA classifier proves the efficiency of both the techniques with a better recognition rate for ELSDSR database.

Second, the good recognition score is obtained using 60 Mel-spaced filterbanks and higher dimensional (52) MFCC features with LDA classifier for Hindi digit database.

Third, using MFCCs with combination of LPCCs as input features and LDA with combination of PCA as classification technique is better for the recognition of speakers on Hindi digit database. This method not only improves the recognition rate but also reduces the computational cost.

FUTURE WORK

Speaker recognition is a developing field and a lot of research is going on to improve the performance of the system.

In this work MFCCs and LPCCs have been used for the feature extraction technique, but we know that the spectral representation of the speech signal is not robust to acoustical variance like background noise. Thus the performance of the recognizer can be evaluated using a new set of feature extraction technique like that of wavelet transform, as it is known that the wavelet transform uses multi-resolution property by which the different frequencies are analyzed with different resolutions.

In present study, we have taken only the cepstral features. We can also use HMM and other classifiers for speaker recognition. This time, we have performed text-independenty speaker identification. We can also deal with the task of speaker verification in the future work. One more thing that we wish to include is that, except of using text-dependent or text-independent method, we can deal with text-prompted speaker recognition, whereby the user is prompted to enter a sequence of key words that is chosen randomly everytime which the system is used. This case is required in some practical applications and is more better from text-dependent or text-independent methods.

These are the few tasks, which we can include in our future work. Although there is a great scope in this field, it is a challenging task especially for forensic applications. Hence, we wish to perform an easier and straightforward method for this purpose, which is invariant against the voice disguise and other sources of errors.

REFERENCES

1) Harvey Fletcher, "*Speech and Hearing in Communication*," First Published in 1953, D. Van Nostrand Company, Inc. Princeton, New Jersey, New York.
2) Website: http://www.studentpulse.com/a?id=82.
3) Ing. Milan Sigmund, CSc., "Speaker Recognition, Identifying People by their Voices," Habiltation Thesis, Faculty of Electrical Engineering and Computer Science, Institute of Radio Electronics, Brno University of Technology, 2000.
4) Tommi Kinnunen, "Spectral Features for Automatic Text-Independent Speaker Recognition", Licentiate's Thesis, University of Joensuu, Department of Computer Science, Finland, 2003.
5) P. Denes and M. V. Mathews, "Spoken Digit Recognition Using Time-Frequency Pattern Matching", Journal of the Acoustical society of America, Vol.32, No.11, pp. 1450-1455, 1960.
6) S. Pruzansky, "Pattern Matching Procedure for Automatic Talker Recognition", Journal of the Acoustical society of America, Vol.35, No.3, pp. 354-358, 1963.
7) S. Pruzansky and M. V. Mathews, "Talker-Recognition Procedure Based on Analysis of Variance", Journal of the Acoustical society of America, Vol.36, No.11, pp. 2041-2047, 1964.
8) K. P. LI, J. E. Dammann and W. D. Chapman, "Experimental studies in speaker verification using an Adaptive System", Journal of the Acoustical Society of America, Vol.40, No.5, pp. 966-978, 1966.
9) James W. Glenn and Noebert Kleiner, "Speaker Identification Based on Nasal Phonation", the Journal of Acoustical Society of America, Vol.43, No.2, pp. 368-372, 1968.
10) James E. Luck, "Automatic speaker verification using Cepstral measurements ", Journal of the Acoustical Society of America, Vol.46, No.4 (2), pp. 1026-1032, 1969.
11) B. S. Atal, "Automatic Speaker Recognition based on Pitch Contours", Journal of the Acoustical Society of America, Vol.52, No.6(2), pp. 1687-1697, 1972.
12) Jared J. Wolf, "Efficient-Acoustic Parameters for Speaker Recognition", Journal of the Acoustical Society of America, Vol.51, No.6(2), pp. 2044-2056, 1972.

13) B. S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech wave for Automatic Speaker Identification and Verification", Journal of the Acoustical Society of America, Vol.55, No.6, pp. 1304-1312, 1974.

14) Aaron E. Rosenberg and Marvin R. Sambur, "New techniques for Automatic speaker verification", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-23, No.2, pp. 169-176, 1975.

15) John D. Markel, Beatrice T. Oshika and Augustine H. Gray, "Long Term Feature Averaging for Speaker Recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.ASSP-25, No.4, pp. 330-337, 1977.

16) George R. Dodding, "Speaker Recognition-Identifying People by their voices", Proceedings of the IEEE, Vol.73, No.11, pp. 1651-1664, 1985.

17) Sadaoki Furui, "Speaker-Independent Isolated Word Recognition using Dynamic Features of Speech Spectrum", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.ASSP-34, No.1, pp. 52-59, 1986.

18) Douglas A. Reynolds, "Experimental Evaluation of Features for Robust Speaker Identification", IEEE Transactions on Speech and Audio Processing, Vol.2, No.4, pp. 639-643, 1994.

19) Philippe Thevenaz and Heinz Hugli, "Usefulness of the LPC-residue in Text-independent Speaker Verification", Speech Communication, Vol.17, pp. 145-157, 1995.

20) Chai Wutiwiwatchai, Varin Achariyakulporn, and Chularat Tanprasert, "Text Independent Speaker Identification using LPC and DTW for Thai Language", IEEE TENCON, 1999.

21) Ehab F. M. F. Bardan, Hany Selim, "Speaker Recognition using Artificial Neural Networks Based on Vowel Phonemes", Proceedings of IEEE ICSP, Vol.2, pp. 796-802, 2000.

22) Macros Faundez-Zanuy, Daniel Rodriguzez-Porcheron, "Speaker Recognition Using Residual Signal of Linear and Non-linear Prediction Models", Escola Universtaria Politecnica de Catalunya, Mataro (Barcelona) Spain, 2002.

23) Yang Zhen, Li Canwei, "A New Feature Extraction Based on the Reliability of Speech in Speaker Recognition", IEEE ICSP Proceedings, Vol.1, pp. 536-539, 2002.

24) Wan-Chen Chen, Ching-Tang Heish, and Eugene Lai, "Multiband Approach to Robust Text Independent Speaker Identification", Computational Linguistics and Chinese Language Proceesing, Vol.9, No.2, pp. 63-76, 2004.

25) WU Zunjing, CAO Zhigang, "Improved MFCC Based Feature for Robust Speaker Identification", Tsinghua Science and Technology, Vol.10, No.2, pp. 158-161, 2005.

26) Hanwu Sun, Bin Ma and Haizhou Li, "An Efficient Feature Selection Method for Speaker Recognition", IEEE ICSLP, pp. 1-4, 2008.

27) Yang Hong-Wu, Liu Ya-Li, and Huang De-Zhi, "Speaker Recognition Based on Weighted Mel-Cepstrum", IEEE Forth International Conference on Computer Sciences and Convergence Information Technology, pp. 200-203, 2009.

28) S. Malik and Fayyaz A. Afsar, "Wavelet Transform Based Automatic Speaker Recognition, IEEE 13th International Conference, pp.1-4, 2009.

29) Xugang Lu, and Jianwu Dang, "An investigation of Dependencies between Frequency Components and Speaker Characteristics for Text-independent Speaker Identification", Speech Communication, Vol.50, No.4, pp. 312-322, 2008.

30) Hassen Seddik, Amel Rahmouni and Mounir Sayadi, "Text Independent Speaker Recognition Using the Mel Frequency Cepstral Coefficients and Neural Network Classifier", IEEE, pp. 631-634, 2004.

31) Joseph P. Campell, Jr., Senior member, "Speaker Recognition: A Tutorial", Proceedings of IEEE, Vol.85, No.9, 1997.

32) John Holmes and Wendy Holmes, "*Speech Synthesis and Recognition*", 2nd edition, Taylor & Francis, Inc. Bristol, PA, USA.

33) J. Ortega-Garcia, J. Gonzalez-Rodriguez and S. Criz-Llanas, "Speech Variability in Automatic Speaker Recognition Systems for Commercial and Forensic Purposes", IEEE AES systems Magazine, Vol.15, No.11, pp. 27-32, 2000.

34) Sadaoki Furui, "Speaker Recognition", Scholarpedia, Vol.3, No.4, pp. 3715, 2008.

35) L. Rabiner and B. H. Juang, "*Fundamentals of Speech Recognition*", 1st ed., Pearson Education, Delhi, 2003.

36) R. P. Sharma, "Recognition of Hindi (Stop) Consonants", Ph.D. thesis submitted to Aligarh Muslim University, Aligarh, India, 2008.

37) J. R. Deller, J. H. L. Hansen and J. G. Proakis, "*Discrete Time Proceesing of Speech Signals*" Piscataway (N.J.), IEEE Press, 2000.

38) I. Khan, "Statistical Study of Hindi Speech Sounds", Ph.D. thesis submitted to Aligarh Muslim University, Aligarh, India, 1990.

39) Omar Farooq, I. Khan, R. P. Sharma, and Preeti Verma, "Classification of Unaspirated Hindi Stop Consonants in Final Position of VC Syllables", National Symposium on Acoustics (NSA), 2010.

40) Shung-Yung Lung, "*Speaker Recognition*", National University of Taiwan, Taiwan.

41) Herbert Gish and Michael Schmidt, "Text Independent Speaker Identification", IEEE Signal Processing Magazine, Vol.11, No.4, pp. 18-31, 1994.

42) Minh N. Do, "An Automatic Speaker Recognition system", Audio Visual Communication Laboratory, Swiss Federal Institute of Technology, Lausanne, Switzerland.

43) Goutam Saha, Ulla S. Yadhunandan, "Modified Mel Frequency Cepstral Coefficient", Department of Electronics and electrical Communication Engineering, Indian Institute of Technology, Kharagpur.

44) Haamid Manzoor Gazi, "Isolated Digit Recognition Using Hidden Markov Model (HMM)", M. Tech. dissertation submitted to Aligarh Muslim University, Aligarh, India, 2005.

45) R. O. Duda, P. E. Hart and G. Stork, "*Pattern Classification*", 2nd ed., John Willy & Sons Press, New York, 2001.

46) Website: http://www.dtreg.com/lda.htm.

47) Jyoti Garg, "Recognition of Emotion from Speech Signal", M. Phil dissertation submitted to Aligarh Muslim University, Aligarh, India, 2006.

48) Website: http://www.imm.dtu.dk/~lf/ELSDSR.htm.

49) Ling Feng and Lars Kai Hansen, "A New Database for Speaker Recognition", Informatics and Mathmatical Modelling, Technical University of Denmark, Denmark, 2004.

50) Abdulnasir aahossen and Said Al-Rawahi, "A Text Independent Speaker Identification System Based on the Zak Transform", Signal Processing on International Journal (SPIJ), Vol.4, No.2, pp. 68-74, 2010.

51) O. Farooq and S. Datta, "Mel Filter-Like Admissible Wavelet Packet Structure for Speech Recognition", IEEE Signal Processing Letters, Vol.8, No.7, pp. 196-198, 2001.

52) Jian-Da Wu, and Bing-Fu Lin, "Speaker Identification Using Discrete Wavelet Packet Transform Technique with Irregular Decomposition", Speech Communication, Vol.36, No.2 (2), pp. 3136-3143, 2009.

APPENDIX – A

List (I):

1. Chicken Little was in the woods one day when an acorn fell on her head. It scared her so much she trembled all over. The poor girl shook so hard, half her feathers fell out.
2. Billions of black, shrimp-size bugs with transparent wings and beady red eyes are beginning to carpet trees, buildings, poles, and just about anything else vertical in the U.S. from the eastern seaboard west through Indiana and south to Tennessee.
3. Oymyakon, in Siberia, is the coldest permanently inhabited place on Earth. Now geographer and adventurer Nick Middleton reveals the locals' secrets for coping with the cold.
4. Few shores are immune from the tide of plastic soda bottles, bags, cartons, and other trash floating on the ocean today. Now a new study suggests the problem runs deeper: Microscopic bits of plastic permeate the world's beaches and marine environment.
5. One hundred years later, the life of the Negro is still sadly crippled by the manacles of segregation and the chains of discrimination.
6. People are finding medieval toys in Britain's Thames River-and these toys have been changing how historians view the lives of medieval kids.
7. My friend Tricia suggests me to drive the woods to watch the poor bear being hunted for pleasure. And I say yes.

List (II):

(These sentences are extracted from NOVA Home: http:// www.pbs.org/wgbh/nova/pyramid/explore/)

1. There are days when the sand blows ceaselessly, blanketing the remains of a powerful dynasty that ruled Egypt 5,000 years ago.
2. When the wind dies down and the sands are still, a long shadow casts a wedge of darkness across the Sahara, creeping ever longer as the North African sun sinks beyond the horizon.
3. Five thousand years ago, the fourth dynasty of Egypt's Old Kingdom was a highly advanced civilization where the kings, known as pharaohs, were believed to be gods.
4. They lived amidst palaces and temples built to honor them and their deified ancestors.
5. "Pharaoh" originally meant "great house", but later came to mean king.
6. This web site will show you science in action – bringing you face to face with the evidence archaeologists use to understand the meaning of Giza's pyramids, and to the process of evaluating the finds they will uncover beneath the sands of the plateau.
7. Before looking closely at pharaonic society and the beginning of the Pyramid Age, one first has to step into Egypt's landscape and take a look around.
8. Ancient Egyptians called their land "Kemet", which meant "black", after the black fertile silt-layered soil that was left behind each year during the annual inundation, when the Nile flooded the fields.
9. The most prevalent color of the desert, however, is a decidedly reddish-yellow ochre.
10. The Egyptians called the desert "deshret", meaning "red", and this endless carpet of sand covers an estimated 95% of Egypt, interrupted only by the narrow band of green carved by the waters of the Nile.

11. It was at this time that hieroglyphic writing made its first appearance, in the tombs and treasures of the pharaohs.
12. To seal the unification of Upper and Lower Egypt, Menes founded the capital city of the kingdom at the place where the two met: at the apex of the Nile, where it fans out onto the fertile silt plain.
13. The fortress city was named "White Walls" by Menes, but it is known today by its Greek name, Memphis.
14. For much of the 3,000 years of ancient Egypt, it remained the capital seat of the pharaohs.
15. Only 20 miles to the north of Memphis is the modern capitol, Cairo, still situated near the juncture of the Nile valley and the delta.
16. How does the pyramid fit into early Egyptian life?
17. Pyramids today stand as a reminder of the ancient Egyptian glorification of life after death, and in fact, the pyramids were built as monuments to house the tombs of the pharaohs.
18. Death was seen as merely the beginning of a journey to the other world.
19. In this society, each individual's eternal life was dependent on the continued existence of their king, a belief that made the pharaoh's tomb the concern of the entire kingdom.
20. Pictures on the walls of tombs tell us about the lives of the Kings and their families.
21. We know pyramids were built during a king's lifetime because hieroglyphs on tomb walls have been found depicting the names of the gangs who built the pyramids for their kings.
22. Furniture and riches were buried with the king so he would have the forts of his familiar comforts of his lifetime buried near him.
23. Whole subdivisions of tombs of those in high positions in the court of a king can be found surrounding the pyramids of Giza.

24. These are primarily mastabas, or covered rectangular tombs that consist of a deep burial shaft, made of mud brick and half-buried by the drifts of sand on the plateau.
25. The first pyramid was the Step Pyramid at Saqqara, built for King Zoser in 2750 BC.
26. This first application of large scale technology, however, is often attributed to Imhotep, the architect of the Step Pyramid.
27. He was not a pharaoh, but was the Director of Works of Upper and Lower Egypt.
28. The superstructure of the pyramid was made of small limestone blocks and desert clay.
29. Inside, the burial chamber and chamber and storage spaces for Zoser's gave goods were carved out of the earth and rock beneath the structure.
30. Imhotep's intent was to mimic the basic structure of King Zoser's palatial home in the burial chamber.
31. The tomb, like those that followed, was meant to be a replica of the royal place.
32. In early tombs, the central area was always the burial place.
33. It is thought that in 816 AD Caliph al-Mamun first ordered workers to blast through the blocked stone entrance in order to explore within Khufu's pyramid.
34. But looters, probably from dynastic Egyptian times, had already absconded with King Khufu's burial treasures and his body.
35. This is true of all of the pyramids at Giza, so very little is known about Khufu or any of his successors who were buried at Giza.
36. Archaeologists, nonetheless, continue to look for pieces of this puzzle to further our understanding of the pyramid Age and the Pharaohs that ruled Egypt.
37. Standing at the base of the Great Pyramid, it is hard to imagine that this monument—which remained the tallest building in the world until early in this century—was built in just under 30 years.
38. It presides over the plateau of Giza, on the outskirts of Cairo, and is the last survivor of the Seven Wonders of the World.

39. Today, Giza is a suburb of rapidly growing Cairo, the largest city in Africa and the fifth largest in the world.
40. About 2,550 B. C., King Khufu, the second pharaoh of the fourth dynasty, commissioned the building of his tomb at Giza.
41. Some Egyptologists believe it took 10 years just to build the ramp that leads from the Nile valley floor to the pyramid, and 20 years to construct the pyramid itself.
42. On average, the over two million blocks of stone used to build Khufu's pyramid weigh 2.5 tons, and the heaviest blocks, used as the ceiling of Khufu's burial chamber, weigh in at an estimated 40 to 60 tons.
43. This question has long been debated, but many Egyptologists agree the stones were hauled up ramps using ropes of papyrus twine.
44. The popular belief is that the gradually sloping ramps, build out of mud, stone and wood were used as transportation causeways for moving the large stones to their positions up and around the four sides of the pyramids.
45. Giza, however, is more than just three pyramids and the Sphinx.
46. Each pyramid has a mortuary temple and a valley temple linked by long causeways that were roofed and walled.

APPENDIX – B

Table 4.3: Confusion matrix obtained for ELSDSR database using MFCCs features with LDA classifier.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
A	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0
J	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0
O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0
U	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100

Table 4.4: Confusion matrix of Hindi Digit Database (HDD) obtained using 52-dimensional MFCCs with LDA classifier.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
A	**80**	0	5	0	0	0	0	0	5	0	5	0	0	0	0	0	0	0	5	0	0
B	0	**90**	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0
C	10	0	**75**	0	0	0	0	0	0	0	0	0	0	5	0	0	5	0	5	0	0
D	0	0	0	**90**	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0
E	0	0	5	0	**95**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	15	20	0	**55**	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0
G	0	0	0	0	5	0	**95**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	5	0	**95**	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0	0
J	0	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0
K	0	30	0	0	0	0	0	0	0	0	**70**	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	**90**	0	5	0	5	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	5	**95**	0	0	0	0	0	0	0	0
N	10	0	15	0	0	0	0	5	0	0	0	0	20	**50**	0	0	0	0	0	0	0
O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0
P	10	0	25	0	0	0	0	5	0	0	0	0	0	25	0	**30**	0	0	5	0	0
Q	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	**80**	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	**85**	0	0	0
S	10	0	35	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	**45**	0	0
T	0	5	0	0	0	0	0	0	0	0	5	0	5	0	0	0	0	0	0	**85**	0
U	5	0	0	0	0	**0**	0	0	0	0	0	0	0	5	5	0	0	0	5	0	**80**

Table 5.1: Confusion matrix of HDD obtained using MFCCs and LPCCs with LDA classifier.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
A	**75**	5	5	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	5	0	0
B	0	**90**	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0
C	20	0	**75**	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0
D	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	5	0	**95**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	5	60	0	**35**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	**95**	0	0	0	0	0	0	5	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0	0
J	0	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0
K	0	30	0	0	0	0	0	0	0	0	**70**	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	**90**	0	10	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0
N	0	0	20	0	0	0	0	5	0	0	5	0	15	**35**	0	0	0	0	20	0	0
O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0
P	5	0	0	0	0	0	0	0	0	0	0	0	5	20	0	**55**	0	0	15	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**100**	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	**95**	0	0	0
S	5	0	10	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	**60**	0	0
T	0	5	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	**85**	0
U	0	0	0	0	0	0	0	0	0	0	0	0	0	15	5	0	0	0	10	0	**70**

Table 5.2: Confusion matrix of HDD obtained using MFCC features with NNet classifier.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
A	**35**	0	40	0	0	0	0	0	0	0	5	0	5	5	0	5	0	0	5	0	0
B	0	**80**	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0
C	0	0	**85**	0	0	0	0	5	0	5	0	0	0	0	0	0	0	0	5	0	0
D	0	0	0	**95**	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	**90**	0	5	0	0	0	0	0	0	0	5	0	0	0	0	0	0
F	0	0	10	0	0	**75**	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0
G	0	0	5	0	10	0	**85**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	20	0	0	0	**80**	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0	0
J	0	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0
K	0	30	0	0	0	0	0	0	0	0	**70**	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	**90**	0	5	0	0	0	5	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0
N	15	0	45	0	0	0	0	15	0	0	10	0	15	**0**	0	0	0	0	0	0	0
O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0
P	5	0	40	0	0	0	5	5	0	0	0	0	0	0	0	**45**	0	0	0	0	0
Q	0	0	25	0	0	30	5	0	0	0	0	0	0	0	0	5	**35**	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	**90**	0	0	0
S	5	0	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**65**	0	0
T	0	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**60**	0
U	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	**90**

Table 5.3: Confusion matrix of HDD obtained using MFCCs and LPCCs features with NNet classifier.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
A	**60**	5	15	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	15	0	0
B	0	**90**	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0
C	20	0	**70**	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0
D	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	**95**	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	5	0	10	5	0	**80**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	**90**	0	0	0	0	0	10	0	0	0	0	0	0	0
I	0	0	10	0	0	0	0	0	**90**	0	0	0	0	0	0	0	0	0	0	0	0
J	0	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0
K	0	45	0	0	0	0	0	0	0	0	**55**	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	**95**	0	0	5	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	5	0	**90**	5	0	0	0	0	0	0	0
N	5	0	20	0	0	0	0	5	0	0	15	0	10	**30**	0	15	0	0	0	0	0
O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0
P	0	0	10	0	0	0	0	5	0	0	0	0	0	10	0	**30**	0	0	45	0	0
Q	0	0	0	5	0	0	5	0	0	0	0	0	0	0	0	0	**90**	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	0	30	0	0	0	0	**70**	0	0	0
S	15	0	5	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	**75**	0	0
T	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**95**	0
U	5	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0	**80**

Table 5.4: Confusion matrix obtained using 52-dimensional MFCCs and PCA with LDA classifier.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
A	**70**	0	5	0	0	0	0	0	5	0	0	0	0	5	0	0	5	0	5	5	0
B	0	**90**	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0
C	20	0	**75**	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	5	0	0
D	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	10	0	**90**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	15	0	**80**	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0
G	0	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	5	0	**95**	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0	0
J	0	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0
K	0	25	0	0	0	0	0	0	0	0	**75**	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	5	0	0	0	**90**	0	5	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0
N	5	0	15	0	0	0	0	15	0	0	0	0	20	**45**	0	0	0	0	0	0	0
O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0
P	10	0	20	0	0	0	0	5	0	0	0	0	0	25	0	**30**	0	0	10	0	0
Q	0	0	40	0	0	0	0	0	0	0	0	0	0	0	0	0	**60**	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	**85**	0	0	0
S	15	0	10	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	**55**	0	0
T	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**85**	0
U	10	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	5	0	**80**

Table 5.5: Confusion matrix of HDD obtained using MFCCs + LPCCs and PCA with LDA classifier.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
A	**70**	5	5	0	0	0	0	0	0	5	0	0	0	5	0	10	0	0	0	0	0
B	0	**90**	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0
C	10	0	**80**	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	**95**	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0
F	0	0	0	30	0	**70**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	**95**	0	0	0	0	0	0	5	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0	0
J	0	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0
K	0	20	0	0	0	0	0	0	0	0	**80**	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	**90**	0	10	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0
N	5	0	15	0	0	0	0	20	0	0	0	0	15	**30**	0	0	0	0	10	5	0
O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0
P	10	0	0	0	0	0	0	0	0	0	0	0	5	20	0	**45**	0	0	20	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**100**	0	0	0	0
R	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	**90**	0	0	0
S	10	0	5	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	**60**	0	0
T	0	5	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	**85**	0
U	5	0	0	0	0	0	0	0	0	0	0	0	0	5	5	0	0	0	5	0	**80**

APPENDIX – C

MATLAB program for Linear Discriminant Classifier.

(Designed for ELSDSR database)

```
clear all
          clc

          func='linear';          %discriminant function.
          [a]=wk1read('filename.wk1');
          A=a(1:9,:);             %all the samples of speaker 1.
          B=a(10:18,:);           %all the samples of speaker 2.
          C=a(19:27,:);           %all the samples of speaker 3.
          D=a(28:36,:);           %all the samples of speaker 4.
          E=a(37:45,:);           %all the samples of speaker 5.
          F=a(46:54,:);           %all the samples of speaker 6.
          G=a(55:63,:);           %all the samples of speaker 7.
          H=a(64:72,:);           %all the samples of speaker 8.
          I=a(73:81,:);           %all the samples of speaker 9.
          J=a(82:90,:);           %all the samples of speaker 10.
          K=a(91:99,:);           %all the samples of speaker 11.
          L=a(100:108,:);         %all the samples of speaker 12.
          M=a(109:117,:);         %all the samples of speaker 13.
          N=a(118:126,:);         %all the samples of speaker 14.
          O=a(127:135,:);         %all the samples of speaker 15.
          P=a(136:144,:);         %all the samples of speaker 16.
          Q=a(145:153,:);         %all the samples of speaker 17.
          R=a(154:162,:);         %all the samples of speaker 18.
          S=a(163:171,:);         %all the samples of speaker 19.
          T=a(172:180,:);         %all the samples of speaker 20.
          U=a(181:189,:);         %all the samples of speaker 21.
          V=a(190:198,:);         %all the samples of speaker 22.
          W=a(199:207,:);         %all the samples of speaker 23.


          A1=A(1:7,:);            %training samples of speaker 1.
          A2=A(8:9,:);            %testing samples of speaker 1.
          B1=B(1:7,:);            %training samples of speaker 2.
          B2=B(8:9,:);            %testing samples of speaker 2.
          C1=C(1:7,:);            %training samples of speaker 3.
          C2=C(8:9,:);            %testing samples of speaker 3.
          D1=D(1:7,:);            %training samples of speaker 4.
```

```
D2=D(8:9,:);            %testing samples of speaker 4.
E1=E(1:7,:);            %training samples of speaker 5.
E2=E(8:9,:);            %testing samples of speaker 5.
F1=F(1:7,:);            %training samples of speaker 6.
F2=F(8:9,:);            %testing samples of speaker 6.
G1=G(1:7,:);            %training samples of speaker 7.
G2=G(8:9,:);            %testing samples of speaker 7.
H1=H(1:7,:);            %training samples of speaker 8.
H2=H(8:9,:);            %testing samples of speaker 8.
I1=I(1:7,:);            %training samples of speaker 9.
I2=I(8:9,:);            %testing samples of speaker 9.
J1=J(1:7,:);            %training samples of speaker 10.
J2=J(8:9,:);            %testing samples of speaker 10.
K1=K(1:7,:);            %training samples of speaker 11.
K2=K(8:9,:);            %testing samples of speaker 11.
L1=L(1:7,:);            %training samples of speaker 12.
L2=L(8:9,:);            %testing samples of speaker 12.
M1=M(1:7,:);            %training samples of speaker 13.
M2=M(8:9,:);            %testing samples of speaker 13.
N1=N(1:7,:);            %training samples of speaker 14.
N2=N(8:9,:);            %testing samples of speaker 14.
O1=O(1:7,:);            %training samples of speaker 15.
O2=O(8:9,:);            %testing samples of speaker 15.
P1=P(1:7,:);            %training samples of speaker 16.
P2=P(8:9,:);            %testing samples of speaker 16.
Q1=Q(1:7,:);            %training samples of speaker 17.
Q2=Q(8:9,:);            %testing samples of speaker 17.
R1=R(1:7,:);            %training samples of speaker 18.
R2=R(8:9,:);            %testing samples of speaker 18.
S1=S(1:7,:);            %training samples of speaker 19.
S2=S(8:9,:);            %testing samples of speaker 19.
T1=T(1:7,:);            %training samples of speaker 20.
T2=T(8:9,:);            %testing samples of speaker 20.
U1=U(1:7,:);            %training samples of speaker 21.
U2=U(8:9,:);            %testing samples of speaker 21.
V1=V(1:7,:);            %training samples of speaker 22.
V2=V(8:9,:);            %testing samples of speaker 22.
W1=W(1:7,:);            %training samples of speaker 23.
W2=W(8:9,:);            %testing samples of speaker 23.

z=7;
gr1=ones(z,1);          %class 1 representing speaker 1.
gr2=2*ones(z,1);        %class 2 representing speaker 2.
```

```
gr3=3*ones(z,1);        %class 3 representing speaker 3.
gr4=4*ones(z,1);        %class 4 representing speaker 4.
gr5=5*ones(z,1);        %class 5 representing speaker 5.
gr6=6*ones(z,1);        %class 6 representing speaker 6.
gr7=7*ones(z,1);        %class 7 representing speaker 7.
gr8=8*ones(z,1);        %class 8 representing speaker 8.
gr9=9*ones(z,1);        %class 9 representing speaker 9.
gr10=10*ones(z,1);      %class 10 representing speaker 10.
gr11=11*ones(z,1);      %class 11 representing speaker 11.
gr12=12*ones(z,1);      %class 12 representing speaker 12.
gr13=13*ones(z,1);      %class 13 representing speaker 13.
gr14=14*ones(z,1);      %class 14 representing speaker 14.
gr15=15*ones(z,1);      %class 15 representing speaker 15.
gr16=16*ones(z,1);      %class 16 representing speaker 16.
gr17=17*ones(z,1);      %class 17 representing speaker 17.
gr18=18*ones(z,1);      %class 18 representing speaker 18.
gr19=19*ones(z,1);      %class 19 representing speaker 19.
gr20=20*ones(z,1);      %class 20 representing speaker 20.
gr21=21*ones(z,1);      %class 21 representing speaker 21.
gr22=22*ones(z,1);      %class 22 representing speaker 22.
gr23=23*ones(z,1);      %class 23 representing speaker 23.


gr=cat(1,gr1,gr2,gr3,gr4,gr5,gr6,gr7,gr8,gr9,gr10,gr11,gr12,gr13,...
gr14,gr15,gr16,gr17,gr18,gr19,gr20,gr21,gr22,gr23);

%concatenation of training samples of each class.

class_tr=cat(1,A1,B1,C1,D1,E1,F1,G1,H1,I1,J1,K1,L1,M1,N1,O1,P1,Q1,...
R1,S1,T1,U1,V1,W1);

%concatenation of test samples of each class.

class_ts=cat(1,A2,B2,C2,D2,E2,F2,G2,H2,I2,J2,K2,L2,M2,N2,O2,P2,Q2,...
R2,2,T2,U2,V2,W2);

%classifies each sample in the test data into one of the groups in
training using discriminant function.
        p=classify(class_ts,class_tr,gr,func);

%computing error in classifying the samples of test set.
        err=zeros(1,23);
        for i=1:23
```

```
    for j=1:2
        if p(j+2*(i-1))~=i
            err(i)=err(i)+1;
        end
    end
end
err;
Per_err=sum(err)/46*100;

efficiency=(100-Per_err);

out=reshape(p,2,23);
con_mat=zeros(23,23);
con_mat1=zeros(23,23);
for i=1:2
    for j=1:23
        if out(i,j)==1
            con_mat(1,j)=con_mat(1,j)+1;
            con_mat1(1,j)=(con_mat(1,j)/2)*100;
        elseif out(i,j)==2
            con_mat(2,j)=con_mat(2,j)+1;
            con_mat1(2,j)=(con_mat(2,j)/2)*100;
        elseif out(i,j)==3
            con_mat(3,j)=con_mat(3,j)+1;
            con_mat1(3,j)=(con_mat(3,j)/2)*100;
        elseif out(i,j)==4
            con_mat(4,j)=con_mat(4,j)+1;
            con_mat1(4,j)=(con_mat(4,j)/2)*100;
        elseif out(i,j)==5
            con_mat(5,j)=con_mat(5,j)+1;
            con_mat1(5,j)=(con_mat(5,j)/2)*100;
        elseif out(i,j)==6
            con_mat(6,j)=con_mat(6,j)+1;
            con_mat1(6,j)=(con_mat(6,j)/2)*100;
        elseif out(i,j)==7
            con_mat(7,j)=con_mat(7,j)+1;
            con_mat1(7,j)=(con_mat(7,j)/2)*100;
        elseif out(i,j)==8
            con_mat(8,j)=con_mat(8,j)+1;
            con_mat1(8,j)=(con_mat(8,j)/2)*100;
        elseif out(i,j)==9
            con_mat(9,j)=con_mat(9,j)+1;
            con_mat1(9,j)=(con_mat(9,j)/2)*100;
```

```
        elseif out(i,j)==10
            con_mat(10,j)=con_mat(10,j)+1;
            con_mat1(10,j)=(con_mat(10,j)/2)*100;
        elseif out(i,j)==11
            con_mat(11,j)=con_mat(11,j)+1;
            con_mat1(11,j)=(con_mat(11,j)/2)*100;
        elseif out(i,j)==12
            con_mat(12,j)=con_mat(12,j)+1;
            con_mat1(12,j)=(con_mat(12,j)/2)*100;
        elseif out(i,j)==13
            con_mat(13,j)=con_mat(13,j)+1;
            con_mat1(13,j)=(con_mat(13,j)/2)*100;
        elseif out(i,j)==14
            con_mat(14,j)=con_mat(14,j)+1;
            con_mat1(14,j)=(con_mat(14,j)/2)*100;
        elseif out(i,j)==15
            con_mat(15,j)=con_mat(15,j)+1;
            con_mat1(15,j)=(con_mat(15,j)/2)*100;
        elseif out(i,j)==16
            con_mat(16,j)=con_mat(16,j)+1;
            con_mat1(16,j)=(con_mat(16,j)/2)*100;
        elseif out(i,j)==17
            con_mat(17,j)=con_mat(17,j)+1;
            con_mat1(17,j)=(con_mat(17,j)/2)*100;
        elseif out(i,j)==18
            con_mat(18,j)=con_mat(18,j)+1;
            con_mat1(18,j)=(con_mat(18,j)/2)*100;
        elseif out(i,j)==19
            con_mat(19,j)=con_mat(19,j)+1;
            con_mat1(19,j)=(con_mat(19,j)/2)*100;
        elseif out(i,j)==20
            con_mat(20,j)=con_mat(20,j)+1;
            con_mat1(20,j)=(con_mat(20,j)/2)*100;
        elseif out(i,j)==21
            con_mat(21,j)=con_mat(21,j)+1;
            con_mat1(21,j)=(con_mat(21,j)/2)*100;
        elseif out(i,j)==22
            con_mat(22,j)=con_mat(22,j)+1;
            con_mat1(22,j)=(con_mat(22,j)/2)*100;
        elseif out(i,j)==23
            con_mat(23,j)=con_mat(23,j)+1;
            con_mat1(23,j)=(con_mat(23,j)/2)*100;
        end
    end
```

```
            end

%confusion matrix of the order of 23 x 23 for the recognition of 2
samples of test data of 23 speakers.
            con_mat=con_mat';

%confusion matrix of the order of 23 x 23 for the average recognition
of all the test samples of 23 speakers.
            con_mat1=con_mat1';
```

MATLAB program for Principal Component Analysis.

```
clear all
clc

[b]=wk1read('filename.wk1');
features=b;
N=53;                          % Number of principal components

% pca_kpm Compute top N principal components using eigs or svd.
[pc_vec]=pca_kpm(features,N, 'eigs');
```